



SCALE DEVELOPMENT

量表编制

理论与应用

■（美）罗伯特·F·德威利斯 著

第2版

■ 魏勇刚 龙长权 宋 武 译
■ 李 红 审校

重庆大学出版社

Authorized translation from the English language edition, entitled SCALE DEVELOPMENT: THEORY AND APPLICATION, 2nd edition by Robert F. Devellis, published by Sage Publications, Inc., Copyright 2003 by Sage Publications, Inc.
All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher. CHINESE SIMPLIFIED language edition published by CHONGQING UNIVERSITY PRESS, Copyright 2003 by Chongqing University Press.
量表编制:理论与应用,第2版,作者:罗伯特·F·德威利斯。原书英文版由 Sage 出版公司出版。原书版权属 Sage 出版公司。
本书简体中文版专有出版权由 Sage 出版公司授予重庆大学出版社,未经出版者书面许可,不得以任何形式复制。
版贸渝核字(2003)第10号。

图书在版编目(CIP)数据

量表编制:理论与应用/(美)德威利斯(Devellis,R.F.)著;魏勇刚 龙长权 宋武译——重庆:重庆大学出版社,2004.10
(万卷方法)
ISBN 7-5624-3280-5

Ⅰ.量… Ⅱ.①德…②魏…③龙…④宋… Ⅲ.量表—编制—应用—社会科学—科学研究 Ⅳ.C3

中国版本图书馆 CIP 数据核字(2004)第100786号

量表编制:理论与应用
(第2版)
罗伯特·F·德威利斯 著
魏勇刚 龙长权 宋武 译
责任编辑:雷少波 版式设计:周 晓
责任校对:李定群 责任印制:秦 梅
*
重庆大学出版社出版发行
出版人:张鸽盛
社址:重庆市沙坪坝正街174号重庆大学(A区)内
邮编:400030
电话:(023) 65102378 65105781
传真:(023) 65103686 65105565
网址:http://www.cqup.com.cn
邮箱:fxk@cqup.com.cn(市场营销部)
全国新华书店经销
重庆科情印务有限公司印刷
*
开本:890×1240 1/32 印张:6.125 字数:169千 插页:2页
2004年10月第1版 2004年10月第1次印刷
印数:1—4 000
ISBN 7-5624-3280-5 定价:15.00元

本书如有印刷、装订等质量问题,本社负责调换
版权所有 翻印必究

作译者简介

罗伯特·F·德威利斯　北卡罗莱纳大学(UNC)健康行为与健康教育系(公共健康学院)和心理系(艺术与科学学院)教授。另外，他是UNC的罗伯特伍德约翰逊临床学者计划(医学院)的核心研究成员。德威利斯博士也是UNC老龄化控制中心的测量与方法基础课程的主任以及UNC关节炎多学科临床研究中心的副主任，UNC关节炎多学科临床研究中心的方法学基础课程组的成员。他在美国心理协会健康心理学分会(38)、关节炎基金会的临床／诊断结果／疗法研究部的董事会供职，也在老兵事务部MEI分部任咨询董事。他是《关节炎护理与研究》和《健康教育研究》的编委会成员，以及二十几种杂志的客座编辑、助理编辑和评论员。他目前的研究兴趣包括：研究促进对慢性疾病适应的个体内部因素，以及测量与健康和疾病相联系的社会变量和行为变量。1980年以后，他是一系列由联邦政府和私人基金资助的研究项目的主要研究者或合作研究者。

李　红　西南师范大学心理学院教授，博士生导师。

魏勇刚　心理学硕士，重庆师范大学学前教育学院教师。

龙长权　西南师范大学心理学院心理学硕士研究生。

宋　武　西南师范大学教师。

为方法理性鼓与呼*

——万卷方法策划报告暨出版说明

究竟是中国内地学界对于研究方法的漠视,导致研究方法出版物的匮乏?
还是研究方法出版物的匮乏,导致学界没有对方法问题予以应有的重视?
这是一个鸡生蛋蛋生鸡的问题。
作为图书出版的编辑人和策划人,对之多论无益。
但,作为图书出版的编辑人和策划人,我们却非常乐意——为方法理性鼓与呼!

我们乐于,也更善于从出版物的数量和质量的比较中,来了解和表述某一类知识的生产和传播情况,以此作为我们图书出版策划的基础。同样,在万卷方法的策划之前,我们便对美国、中国台湾及中国内地三者,1999—2002 年间关于社会科学研究方法的图书出版情况做了一个简单的比较:

美国在社会科学研究方法方面出版物的数量之多,至少足以让我们惊讶!由于不便统计,在此,仅就美国的 SAGE 出版公司在社会科学研究方法方面的图书出版情况,做一个简单介绍,以窥其一斑。SAGE 出版公司 4 年中出版的社会科学研究方法方面的书籍(包括再版书)便有两百余种,其中,既有一两百万字的大部头工具书,如 Handbook of Qualitative Research, Handbook of Research Design & Measurement;也有 10 万字左右的口袋书,如仅一个"应用社会科学研究方法系列丛书"(Applied Social Research Meth-

* 本文原载《中华读书报》2004 年 3 月 31 日,第 15 版。在此,根据需要作者对原文做了少量改动。

ods Series）便有 49 个品种，其中每本书对于案例研究方法、调查研究方法、网络调查方法等做了分门别类的介绍。其覆盖面之广，研究之深入、具体、操作性强是我们所难以想象的。如果说，我们与美国的差距尚可理解和接受的话，那么，同一时期，中国台湾在这一方面的研究和传播情况也足以令我们汗颜。4 年间中国台湾出版了大量深入研究社会科学研究方法方面的书，其中不但有本土作者的原创作品，也有从英美等西方国家引进版权的相对比较成熟的社会科学研究方法方面的作品。更为可贵的是，其中有几家出版公司已经注意从社会科学研究方法的体系着眼来组织自己的策划和出版，在这方面的出版有了一定的规模，内容的组织也显得比较成熟：如韦伯文化事业出版社的“研究方法系列丛书”、弘智文化事业公司的“应用性社会科学调查研究方法系列丛书”等都是发展比较成熟、规模比较大的专门传播社会科学研究方法的系列丛书。

而中国内地方面，同期虽然也出版了一批研究方法的书籍，但无论是规模还是研究内容的深入丰富程度都无法与美国及中国台湾两地相提并论。从规模看，这期间中国内地出版的社会科学研究方法类的著作也就四五十本，难望美国同行之项背，仅与中国台湾同期出版规模大致相当。从此类出版物内容的深入丰富程度来看，大陆出版的社会科学研究方法类出版物主要集中在各学科内部，如教育研究方法、心理学研究方法、社会学研究方法、经济学研究方法、体育科学研究方法等，每本书都面面俱到谈调查、谈案例、谈访谈、谈田野、谈观察等，而很少有对某一种方法进行深入研究

的图书，更没有像美国或中国台湾那样形成比较系统的研究社会科学研究方法的系列丛书。在这一领域，这样的图书结构对于应付大学本科生在研究方法方面的入门性需求（即作为教材）还行，但对于迅速培养一大批拥有科学、理性的研究头脑的学者，特别是对于青年学者，如硕士、博士研究生的成长则是远远不够的。其致命的弱点表现在三个方面：其一，一本书铺得太开而无法细化深入，以至于绝大多数学生虽然上了方法的课，却除了记住了几个名词和对一些方法的理论及应用略知一二之外，至于什么情况下选择什么方法最恰当、具体怎么操作、这种方法可能会有哪方面的不足需要加以处理等研究细节问题仍然处于无理性状态。其二，既然我们的方法建设囿于学科内部，而不能对方法进行纵深的开发，那么，反正学科就那么几个，于是方法书大多是低水平重复。比如教育研究方法的书，现在全国使用做教材的至少也有十多本吧，但你若有心思看的话，一本足矣！其三，出版界对研究方法图书这种淡淡的一笔带过的做法，不仅很难形成一种风气，从意识深处提升起大家对于研究方法的关注，而且更加剧了整个学界对于研究方法的漠视，许多学者只有在著书立说之时才想到似乎从“作品结构的完整性”上看应该谈及方法问题。

社会科学界近年流行两个词语：一曰反思，二曰接轨。所谓反思也即是对于学科的理论基础、学科的结构、学科的历史与未来等问题的全面梳理；所谓接轨也即是希望中国的社会科学研究能够融入世界社科研究的大潮中，与国际上的社会科学研究实现彻底

的、清晰的对话。在此，我们以为，无论是反思还是接轨，中国的社会科学界和传播界都必须投入一些精力来关注、研究、教授和传播社会科学研究方法。惟如此，才能在中国形成完善、科学的社会科学研究方法体系和学科群，也才能使对研究方法关注与理性应用在中国社会科学界深入人心、蔚然成风；惟如此，也才能为中国社会科学研究与国际接轨搭建一个平台。

以上种种，促使我们策划出版了万卷方法系列丛书，献给忠实于人文社会科学研究的人们！

雷少波　崔祝

2004 年 6 月于重庆大学出版社

译者前言

“权，然后知轻重；度，然后知长短。物皆然，心为甚。”这不仅表明了对事物的数量差异进行度量的必要性，也表明了对不太容易观察到的心理现象进行度量的可能性。测量个体的心理现象是比较困难的，然而，经过心理测量学家长达百余年的艰苦努力，现在已经有了不少量表能够较准确地测量到复杂且难以观察的心理现象，这对于心理学家研究人的心理现象起到了十分重要的资料收集作用。今天，心理测量量表已经广泛地应用于心理学、社会学、教育学、经济学等学科领域，是对个体的心理和行为进行定量研究的有效手段之一。因此，无论是作为专业的心理学工作者还是作为广泛领域的社会科学工作者，他们在自己的日常研究工作中都广泛地需要作用“恰当的”量表对其所研究的对象进行数量化的度量，然而，他们往往难以找到真正适合其研究目的的量表，这就需要他们根据自己的研究需要自行编制量表。

然而，量表的编制是一项极富挑战性的工作。一方面，量表的编制涉及的知识面比较广泛，不仅包含有关研究对象的专门知识，还包含诸如统计学、测量学等关于量表的基础学科知识。要透彻地理解量表编制的机制与统计原理，对于那些以社会科学为其知识背景的研究者来讲，无疑是一大挑战。另一方面，一个量表要成为有效地测量个体心理和行为的工具，必须要具有好的信度和效度，而如何保证量表的信度和效度对于研究者来讲，也是一个棘手的问题。因此，有一本关于量表编制的著作来指导研究者的学习和工作，这显然是必需的。

大约 14 年前，在我刚刚开始担任大学教师的时候，我就给教育学专业、心理学专业的本科生上了教育与心理测量学课程。我当时就十分迫切地希望能够有一本理论性、操作性和实用性都很好的关于量表制作的教材用于本科生的教学工作。然而，时至今日有关量表编制的专著可谓凤毛麟角。有关量表编制的问题常常主要是包含在有关统计学或测量学的著作之中，这些著作要么过于强调量表编制的统计学原理和测量学原理，要么只注重理论推导而不重视实例说明和直观明了的表达形式，从而使得读者感到晦涩难懂，可操作性差。

《量表编制：理论与应用》一书是美国著名的健康与心理学研究专家罗伯特·F·德威利斯（Robert F.DeVellis）教授根据自己多年从事研究和教学实践而创作的一本关于量表编制的专著。该书自第 1 版问世以来，就在诸如心理学院、公共健康学院、经济学院、教育学院等机构中作为研究课程的教材被广泛采用。其成功之处在于通俗易懂，直观明了，操作性强，具有可读性。主要表现在以下几个方面：①采用图表的形式，将复杂的统计和测量学原理直观明了地展现给读者；②采用类比的方式来代替数学化的术语，使读者感到浅显易懂；③提供大量的实例，使读者感到可操作性强。本书是该书的第 2 版。除保留和改进了第 1 版“用使基本原理清晰明了的方式来传递信息，并使读者能够窥探看起来像‘黑箱’的各种方法”这一通俗易懂的特点外，第 2 版的特点主要是根据社会发展和研究的需要，增加或扩充了大量有价值的内容，主要包括表面效

度、因素分析、项目反应理论、量表编制指南与实践技巧等。因此，本书无论是从内容上来讲，还是从表现手法上来讲，都具有较高的可读性，具有较高的理论价值和实践价值。本书不仅仅适用于心理学领域的研究者和实践者，也适用于所有社会科学领域的研究者与实践者。

2004 年 1 月份，重庆大学出版社雷少波先生专程到学校来邀请我主持翻译这本专著，本来因为工作忙而准备推辞，但是在我翻阅了该书的原著后，我觉得十分有必要将她译介到中国来，使广大的中国社会科学工作者和专业心理学工作者能够受益于这本优秀的量表编制专著，从而帮助更多的初涉社会科学研究和心理学研究工作的人士提高其科研水平，促进中国社会科学研究和心理学研究的整体进步。为此，我让我的研究生魏勇刚、龙长权和青年教师宋武共同承担了本书的翻译工作，最后由我审校。

尽管我们怀着战战兢兢的心情反复对照原著翻译、审校和修改，但由于译者水平限制，我们的翻译工作难免有所疏漏甚至由于理解错误而至误译，敬请读者在阅读过程中提出宝贵的意见和建议，以便我们在本书的译本修订中加以改进。

李　红

2004 年 10 月于西南师范大学心理学院

英文版前言

作为一本介绍测量概念和测量方法的入门教材，本书的第1版得到了广泛的使用。我确信，其成功之处就在于它使复杂的观点变得通俗易懂，这也是我的目标所在。此书出版的一个极其重要的出发点就在于为了帮助各个水平的学生概念性地（conceptually）理解测量问题。在切普山（Chapel Hill）的北卡罗莱纳州大学的公共健康学院，我给本科生开设的量表编制的课程吸引了许多不同背景的学生。在同一学期内，我的学生里既有只学过一门本科统计课程的，也有攻读定量心理学博士（PH.D.in quantitative psychology）的。教授该课程的经验表明，不同水平的学生都从这一以清晰的、概念性的和非数学化的术语形式呈现的材料中获得了益处。尽管公式在此类课程中是必要的，我尽力用使这些公式清晰明了的方式来解释这些概念，它们只不过是合理地简化了运用于数据中的一系列操作。我尽力在第1版中介绍一些已获得显著成功的教学方式，在此修订版中，我也做了此类尝试。本书的重点在于，用使基本原理清晰明了的方式来传递信息，并使读者能够窥探看起来像“黑箱”的各种方法。

此修订版已做了大量的修改。在修订中，我保留了学生们认为最清晰、最有用的方法，增加了自第1版问世以来更受重视的主题。每一章都有修改，有几章的内容已经很充实。增加了三十多本参考书，也保留了许多经典著作，它们在此版中再次被引用。有几章增加了图表以使关键点直观化。在第1章中，我新增了一些例子，阐明了为什么一些变量需要用很多题项来进行有效的评估，而其他的

变量却不需要，并且对不同的题项组合类型进行了更广泛的讨论。第2、3章的内容经过修改已变得更为清楚。为了做些调整，我在第4章增加了关于表面效度的部分；在第5章列出了量表编制的指导方针，并增加了几个对学生有用的实践技巧；第8章从一个更广阔的角度来看测量，并且有所拓展，包括在何处寻找测量工具、高质量的程序如何作为量表编制的基础，以及与不同题项功能相关的一些问题。余下的两章，与前一版相比，改动最大。为了使因素分析过程更加生动、更可理解，第6章因素分析在报告的内容范围上有了相当大的扩充，并完全重写了。我运用了大量图表来说明文字材料。最后，新增加的第7章介绍了一个在第1版中只简要提及的主题——项目反应理论（item response theory，IRT）。我的目的并不在于教给读者关于IRT的非常复杂的操作性知识和正在研究的方法，而是给他们提供一个概念基础，以帮助他们理解在别处碰到的更难的材料。

尽管增加了第7章的内容，本书的重点仍然是经典的测量方法。毫无疑问，随着必要的分析所需要的更易进行数据处理的计算机程序的运用，像IRT这样的理论定会备受欢迎，但经典的方法不会消失。尽管存在某些理论上的缺陷，那些方法在多种情景中运作得出奇地好。它们的基础和运用都很容易理解。在修订版的不同部分，我强调了一些我认为IRT优越于经典方法的几个十分重要的方面。但是，已有的大量研究表明，经典方法仍运行得很好。当IRT处于优势时，经典理论并不会随之变为陈词滥调。二者将由于它们存

在各自的优缺点而相互并存，从而互相补充。许多应用研究者将不会真正需要除了经典测量以外的技术。因此，更让我担心的，不是那些掌握了已有的最新方法的人和没有掌握已有的最新方法的人之间在测量领域内的差距，而是那些掌握了大量测量概念或方法的人与没有掌握任何测量概念或方法的人之间在测量领域内的差距，我希望本书的出版能够帮助读者缩小这一差距。

罗伯特•F•德威利斯

目　录

1　概　论 …………………………………………………… 2

测量概述 …………………………………………………… 4

测量在社会科学中的历史渊源 ……………………………… 5

测量的后续发展 ……………………………………………… 7

测量在社会科学中的作用 …………………………………… 8

总结与展望 …………………………………………………… 15

2　潜在变量 …………………………………………………… 16

结构与测量 …………………………………………………… 17

作为题项值的假设性因素的潜在变量 ……………………… 18

路径图 ………………………………………………………… 19

测量模型的进一步阐述 ……………………………………… 23

平行测试 ……………………………………………………… 23

其他的模型 …………………………………………………… 27

练　习 ………………………………………………………… 29

3　信　度 …………………………………………………… 30

连续题项与二分题项 ………………………………………… 31

内部一致性 …………………………………………………… 31

以量表分数间的相关为基础的信度 ………………………… 43

概括化理论 …………………………………………………… 48

总　结 ………………………………………………………… 51

练　习 ………………………………………………………… 51

4　效　度 …………………………………………………… 54

内容效度 ……………………………………………………… 55

标准—相关效度 ……………………………………………… 56

结构效度 ……………………………………………………… 59

表面效度 ……………………………………………………… 63

练　习 ………………………………………………………… 65

5 量表编制指南 …… 66
步骤 1:清楚地决定你要测量什么 …… 67
步骤 2:建立一个题项库 …… 70
步骤 3:决定测量的模式 …… 78
步骤 4:让专家评价最初的题项库 …… 95
步骤 5:考虑确认题项的包含性 …… 96
步骤 6:在一个试测样本中测试题项 …… 97
步骤 7:求题项的值 …… 100
步骤 8:优化量表长度 …… 107
练　习 …… 111
6 因素分析 …… 112
因素分析概述 …… 114
因素分析的概念描述 …… 119
因素的解释 …… 136
主成分与共同因子 …… 137
验证性因素分析 …… 140
量表编制中因素分析的使用 …… 142
样本大小 …… 147
结　论 …… 148
7 项目反应理论概述 …… 150
项目难度 …… 152
项目区分度 …… 153
假阳性 …… 155
项目特征曲线 …… 157
IRT 的复杂性 …… 160
何时使用 IRT …… 162
结　论 …… 164
8 广泛研究背景下的测量 …… 166
编制量表之前 …… 167
量表施测之后 …… 171
最后的思考 …… 173
参考文献 …… 174

1 概　论

Overview

测量概述

测量在社会科学中的历史渊源

测量的后续发展

测量在社会科学中的作用

总结与展望

在广阔的社会调查领域，测量是一个焦点。以下面的假设情境为例：

- 健康心理学家面临一个普遍的难题：他所需要的测量量表(measurement scale)往往并不存在，而他需要一个能区分个体看医生时他或她想要(want)什么和预期(expect)发生什么这二者之间差异的测量尺度。先前的研究并没有注意到这两种观点的差异之处，也不存在能精确区分这个差异的测量方法。尽管他可以虚构一些能揭示这一差异的题项，但是"虚构"的题项可能没有信度，或者所需的概念缺乏对效度的说明。
- 流行病学家正在对一个国家进行健康调查，获得了大量数据进行二次分析(secondary analysis)。他想调查感知到的心理压力的某些方面和健康状况之间的关系。尽管在最初的调查中并没有包含有关压力测量的题项，但最初试图测量其他变量的几个题项明显包含了与压力相关的内容。那么，把这些题项组织成一个有信度的、有效度的心理压力的量表是可能的。然而，如果这些糟糕的题项组成的是一个糟糕的压力量表，那么研究者可能会得出一个错误的结论。
- 某营销组试图策划一个关于高价婴儿玩具的商业活动，却失败了。群组聚集(focus groups)分析表明，父母的消费决策强烈地受到此类玩具是否对儿童具有明显的教育意义的影响。营销组猜想，对婴儿有着高教育、高职业期望的父母最易受到这类玩具的吸引。因此，营销组想从一个更大的、地理位置更分散的样本范围内估计这些父母的期望。而对另外的群体的研究表明，要得到一个充分大的消费者样本的难度太大了。

在以上任何一个情境中，对特定的实际领域感兴趣的人在研究刚开始的时候都遇到了一个测量问题。他们中没有谁最初对测量本身感兴趣。然而他们中的每一个人在达到主要的研究目的之前都必须找到一个能量化特定现象的方法。在每一个案例中，"现成的"的测量工具要么是不合适的，要么是不能用的。所有的研究者都认识到，如果他们采用随便的测量方法，极可能只会产生一些不精确的数

据。因而,编制他们自己的测量工具似乎是惟一可行的选择。

许多社会科学研究者遭遇了相似的难题。对这类难题通常的反应是依赖于现有的测量工具,或者是假定那些新近编制的“看起来”不错的问卷题项可以用来进行测量。那些糟糕的测量所共有的借口是,对编制可靠有效的测量工具的方法感到困难和不熟悉,以及很难得到一些关于研究主题的有用信息。研究者试图获得量表编制的技巧,这一努力可能导致他们要么获得的是测量专家所提供的一些太深奥的原始材料,要么得到的是一些太通俗反而不便使用的东西。本书讨论了对这些方法的选择和使用。

测量概述

测量是一个基本的科学活动。我们通过观察人类、物体、事件和过程而获得相关的知识。要弄清楚这些观察结果常常需要我们对其量化,即要求我们测量那些我们有科学兴趣的事物。测量过程与其所服务的更广泛的科学问题相互作用,二者间的边界常常是察觉不到的。二者的交互作用常存在于一个实体被探测或被精炼(refined)的时候,或者决定怎样量化一个感兴趣的现象的时候,以及推理给现象本身提供了启示的时候。例如,史密斯、厄普和德维利斯(Smith, Earp & DeVellis, 1995)调查了妇女对受虐(battering)的感受。建立在理论分析基础上的一个概念化的模型显示了有六种不同的感受。旨在编制一个测量这些感受的量表的实验指出,在受虐和未受虐的妇女中,一个很流行的、更简单的概念完整地解释了研究的参与者是怎样对给予的 40 个题项中的 37 项进行回答的。这一发现表明,研究者认为的一个复杂的变量集合实际上被生活在社区中的妇女感受到了。在她们眼中,那不过是一个单一的、广泛的现象罢了。因此,在探测妇女关于受虐感受的过程中,我们发现了关于这些感受结构的新的东西。

邓肯(Ducan, 1984)认为,测量的根基在于社会程序(social processes),这些程序以及它们的测量实际上都先于科学:“所有的测量……都是社会测量。物理测量也是以社会为目的的”(p. 35)。

邓肯注意到，最早的社会测量程序，如投票、人口普查以及工作提升系统等，“最初似乎是为了满足大众的需要，而不仅仅是为了合乎科学好奇心而进行的实验。”(p. 106)他进一步指出，同样的程序“可以从物理学的历史中得出：古代的人在解决社会和实践问题的过程中，成功地实现了对长度或距离、面积、数量、重量和时间的测量，物理科学就是建立在这些成就基础之上的。”(p. 106)

无论最初的动机是什么，科学的每一个领域都发展了自身的一套测量程序。例如，物理学发展了特定的方法和设备来研究亚原子微粒。在社会行为科学领域，心理测量学是作为关注于测量心理和社会现象的一门附属专业而发展起来。具有代表性的是，所用的测量程序都是问卷调查，而变量的性质是一个更广泛的理论框架中的一部分。

测量在社会科学中的历史渊源

早期例子

常识和历史记录支持了邓肯的观点：社会需要使得测量在科学出现以前就得到了发展。毫无疑问，一些测量形式已经成为我们种族自史前时期以来所具有的技能中的一部分。最早的人们必须对物体、财产以及对手做出评估，比如根据对手的某些特点(如体格)来对其做出判断。邓肯(1984)引用圣经上的文字以说明其对测量的关注(例如：A false balance is an abomination to the Lord, but a just weight is a delight. 即一个虚假的天平是对上帝的蔑视，而一个公平的砝码是一种快乐)，并指出亚里士多德的作品中涉及了负责检查重量和测量的官员。阿纳斯塔希(Anastasi, 1968)指出，古希腊时所使用的苏格拉底方法在某种程度上可以被看作是知识测验，它涉及以一种什么样的方式来理解事物。迪布瓦博士(P. H. DuBois)在他1964年的论文中描述到，中国早在公元前2200年就进行了行政事务的测量。赖特(Wright, 1999)引用了古代的关于精确测量的其他的一些重要例子，包括7世纪建立

在“七重”(weight of seven)基础上的穆斯林苛税。他还指出法国革命的爆发在某种程度上是由于农民已经受够了不公正的测量制度而导致的。

统计方法的出现和智力测验的作用

农纳利(Nunnally,1978)指出,尽管系统的观察方法仍在继续进行,但由于没有统计方法的运用,一直阻碍着人类能力测量科学的发展。直到 19 世纪下半期,统计方法才开始被运用。邓肯也发现(1984),在除了几何学以外的大部分数学领域,系统的观察方法在基础的统计方法建立之前(他也认为基础的统计方法的建立是在 19 世纪)已达千年之久。达尔文在进化论上所做的工作以及他的观察和跨物种的系统变量的测量,使得适当的统计方法在 19 世纪最终得以发展。他的堂兄弟高尔顿男爵(Sir Francis Galton)把对差异的系统观察扩展到了人类。高尔顿的主要关注点在于解剖特质和智力特质(anatomical and intellectual traits)的遗传。被称为“统计学的奠基者”(例如,Allen & Yen,1979,P. 3)的卡尔·皮尔逊(Karl Pearson)是高尔顿的一个晚辈同事,他设计了需要用于检查变量间系统关系的数学方法,包括以他名字命名的积矩相关系数(product-moment correlation coefficient)。这使得科学家能够量化变量间相互作用的程度。查尔斯·斯皮尔曼(Charles Spearman)承接其前辈的研究传统,为 20 世纪初因素分析的发展和普及化奠定了基础。值得一提的是,许多早期正式测验的贡献者(包括在 20 世纪初期,在法国发展智力测验的阿尔弗雷德·比纳(Alfred Binet))都对智力测验感兴趣。因此,许多早期测量学的工作都运用在“智力测验”中。

心理物理学(psychophysics)的作用

现代测量学的另一个历史根源来自于心理物理学。把物理学的研究程序用于研究感觉的尝试引起了关于测量本质的长时间的争论。纳仁和卢斯(Narens & Luce)1986 年总结了这一争论。他们指出,19 世纪晚期赫尔曼(Hermann von Helmholtz)发现了像长度和质量这样的物理属性拥有如正实数一样的内部数学结构。

比如时间或长度单位可以像普通数一样排序和添加。20 世纪早期,争论继续进行。英国科技发展协会委员会(The Commission of the British Association for Advancement of Science)认为,心理变量的基本测量因其在排序或添加感官知觉时面临其固有的难题而无法进行。斯蒂芬(S. Smith Stevens)认为,可用于长度或质量的严格添加并非必要,个体可以对声音强度做出还算连续的比率判断。例如,他们可以判断一种声音强度是另一种声音强度的两倍或是一半。这种比率属性使得来自这些测量中的数据可以进行数学处理。斯蒂芬因其把测量分为定类(nominal)、定序(ordinal)、定距(interval)、定比(ratio)这几个尺度而备受关注。他还指出,响度的判断遵循一个比例尺度(邓肯 1984)。就在斯蒂芬提出其心理物理测量等级的合法性时,瑟斯顿(Louis L. Thurstone)正在发展其因素分析的数学基础。瑟斯顿的兴趣横跨心理物理学和智力。斯蒂芬曾称赞瑟斯顿是把心理物理方法运用到社会刺激测量中去的人(邓肯,1984)。因此,他的工作表明,具有不同历史渊源的心理测量理论和心理测量基础有相互融合的趋势。

测量的后续发展

基本概念的发展

斯蒂芬的测量概念如同他本人一样有影响力,但那绝不是最终的定论。他把测量定义为"根据规则对物体和事件进行的数字分配"(邓肯,1984)。邓肯(1984)向这一定义提出了挑战。他认为斯蒂芬的定义正如"弹钢琴时只根据某些模式敲打乐器的键盘"一样,并不完善,测量不仅仅是数字的分配,还应包括遵循某一物体或事件的属性……或品质的不同程度进行的数字分配(P. 126)。纳仁和卢斯(1986)认识到了斯蒂芬最初关于测量概念的局限性,并提出了许多改进意见。尽管如此,他们的工作都强调了斯蒂芬得出的基本观点,即是测量模型而不是英国科技发展协会委员会所认可的测量类型导致了测量方法可运用于物理科学和非物理科

学。在本质上，对基本属性进行测量这些工作，使测量程序在社会科学领域的运用具有科学的合法性。

智力测验的发展

尽管“智力测验”（或者说，现在更通俗的“能力测验”）一直是心理测量学传统的活动领域，但它已不是本书的主要讨论对象。当测量的目的是要测量特征而不是能力时，许多心理测量学分支的进步，包括项目反应理论，都欠通俗，可能较难运用到实践中去。随着时间的推移，在不同的测量背景中如何运用这些方法，会逐渐变得比如何对能力进行评估更加重要，我们将在随后的章节中对之进行讨论。因此，我的重点主要是讨论那些在社会和心理现象而不是能力的测量中所使用的一些“经典”的方法。

心理测量领域的扩展

邓肯(1984)指出，社会科学中的心理测量学的影响超越了它最初对感觉和智力的测量。测量学本身是作为一种方法学的范例出现的。邓肯用了三个例子来说明测量学的影响：①心理测量学对信度和效度的定义被广泛使用；②社会科学研究中因素分析备受欢迎；③运用社会科学方法编制的量表所包含的变量数量远远多于心理测量学最初所关注的变量数量(p. 203)。心理测量的概念以及对各种心理和社会现象进行测量的方法的运用，将会是本书其他部分的讨论对象。

测量在社会科学中的作用

理论与测量的关系

我们试图在社会科学中进行测量的现象常常来自于理论。理论在构建我们需要测量的问题的概念体系方面扮演着一个重要的角色，而且，任何科学领域所测量的东西都来自理论。当亚原子微粒通过测量被确认之前，测量仅仅是个理论建构。但是，心理学与

其他社会科学中的理论与物理学理论是不同的。在社会科学中，科学家倾向于依赖大量的只关注相当小范围现象的理论模型，而在物理科学中，理论学家较少使用数字并且研究更综合性的问题。例如，费斯廷格(Festinger)的社会性比较理论(social comparision theory)只关注人类经验中一个相当狭窄的范围：人们通过与他人的比较来评价自己观点或能力；而物理学家会为建立一个十分统一的理论而继续他们的工作，这一理论会在一个单一的概念框架中包含所有关于本质的基本力量。而且，社会科学也不像物理科学发展的那样成熟，虽然其理论的发展要快得多。测量的不可捉摸性、复杂现象的原因的多重性，以及理论本身的发展，都向社会科学研究者提出了严峻的挑战。因此，牢记测量的程序并认识到它们的优势和缺点是尤其重要的。

研究者对他们所感兴趣的现象、存在于假设建构中的抽象关系以及可利用的定性工具了解的越多，就越有能力去编制可靠的、有效的和可用的量表。其中，对研究的某一特定现象的具体细节的了解，可能是众多需要考虑的问题当中最重要的一个。例如，社会性比较理论包含许多方面，每一方面都意味着需要不同的测量策略。一项研究可能会要求对社会性比较下一个操作性定义，然后用其作为其他更高或更低等级的相对参照标准，而另一项研究可能会要求被试参照“典型的个体”从多个维度进行自我评定。通过不同的测量从不同的方面获得的同一普遍现象(如“社会比较”)的信息可能不会产生相对一致的结论(DeVellis et al. ,1991)。事实上，尽管在描述上使用了相同的变量名称，但评估的却是不同的变量。因此，编制一个最适合于研究问题的量表需要理解理论的精妙之处。

不同的测量方法要求不同的评估策略。比如，从一个器皿里拿出大量的硬币，这是可以直接观察到的。然而，绝大多数社会和行为科学家感兴趣的变量是不能直接观察到的，如信念、动机状态、期望、需要、情感和社会角色认知等。某些不能直接观察的变量是可以通过研究程序测定的，但问卷做不到这一点。例如，尽管认知研究者不能直接观察个体是怎样在其自我图式中建构性别信息的，但他们却能通过回忆程序(recall procedures)推断出个体是

怎样建构其关于自我和性别认知的。然而,在许多情况下,用纸笔测试(paper-and-pencil assessment)以外的其他的方法评估社会科学变量是不可能的,也不实际。当我们对测量的理论建构感兴趣的时候,这种情形虽然并不总是发生,但却经常发生。因此,一个对测量雌雄同体感兴趣的研究者可能会发现,凭借一个精心编制的问卷可以比其他方法更容易得到实验信息。

理论的与非理论的测量

在这里,我们承认,虽然此书的重点在于测量理论的建构,但并非所有的纸笔测试都需要理论建构。比如,性别和年龄可以通过问卷中的自我报告来确定。这两个变量可以成为某个理论模型的成分,也可以仅仅是某项研究中对参与者的部分描述,这取决于研究的实际问题。某些情境,如要求被试采用纸笔测试形式来回答一些问题,例如对医院病人的饮食偏好做出评估,就可以是没有理论根据的。在其他的情形中,一个研究可能以非理论化的形式开始,以明确表达的理论结束。比如说,一个市场研究者可能让父母列出一张清单,以列举他们买给孩子的玩具类型。随后,这个研究者可能会探究这些清单所包含的关系模式。在已观察到的玩具消费模式的基础上,研究者会设计出一个消费行为的模型。其他有关非理论化测量的例子是民意测验。例如,要求人们回答他们所用的香皂的品牌或者他们试图在选举中投谁一票,这些都很少涉及潜在的理论建构的问题。因为研究者的兴趣在于被试的反应本身,而不是假定问卷反映了个体的某些特点。

有时,很难区分理论测量和非理论测量的情境。例如,通过探讨投票者对总统候选人的偏爱程度来预测某一选举的结果,与要求被试报告他或她的行为目的是等同的。一个调查者可能要求人们回答他们在投票的决策过程中是怎样做到不是从兴趣出发,而仅仅是从所期望的最后的投票结果来投票的。但是,如果同样的问题出现在测量对特定问题的态度是怎样影响投票者对候选人的偏爱时,那么这一研究就可能隐含着一个阐述精确的理论。在这一情境中,获得投票信息的目的不是为了揭示被试将会怎样投票,而是为了对个体的特征有一个清晰的呈现。在这两种情况中,与

测量理论有关还是无关,涉及关于调查者的意图问题,而不是所用的程序问题。主要对测量理论的建构感兴趣的读者可参考其他作者的著作,如肯维斯和普莱斯(Converse & Presser, 1986)、查迦和布莱尔(Czaja & Blair, 1996)、迪尔曼(Dillman, 2000)、芬克(Fink,1995)、福勒(Fowler,1993,1995)以及韦斯伯格、克劳里克和博文(Weisberg,Krosnick,& Bowen ,1996)。

测量量表

由很多题项构成,并且这些题项构成一个复合分数,试图揭示不能轻易用直接方法来观察的理论变量的水平,这样的测量工具常常被称为量表。当我们想测量那些凭借我们对世界的理论理解而相信其存在但又无法直接感知的现象时,我们就编制量表。例如,我们可能会用消沉或焦虑来解释我们所观测到的行为。绝大多数的理论家都同意消沉或焦虑与我们所看到的行为并不等同,但却隐含着某一行为。理论认为,这些现象存在并影响着行为,但它们是无形的。有时,通过它们的行为结果来推测其存在可能是合适的。然而,在其他情况下,我们可能没有办法得到关于行为的信息(如,当我们只能用邮寄的方式来进行调查时),也不能确定怎样解释可得到的行为样本(如,当遭遇某一事件时,绝大多数人会强烈反应,而某一个人却保持消极状态),或者可能不愿意去设想行为与所研究的隐含结构是同构的(如,我们怀疑痛哭是喜悦的结果而不是悲伤的结果)。在那些我们不能把行为作为某一现象来解释的情境中,采用一个建构良好的、有效的量表进行测量是十分有效的。

甚至是从理论中得出的变量,也是一个从相对具体的、可观测的现象到一个相对抽象的、不可观测的现象的一种内在的连续统一体。并不是所有的现象都要求采用多题项量表(multi-item scale)。年龄和性别的确和许多理论有关,但它们却不需要采用多题项变量来进行精确的测量。在很大程度上,这些变量都和具体的、相对清楚的特点(形态学)或事件(出生日期)有关。除非出现某些特殊的情境,如神经受损,否则,被试可以从记忆中轻松地找到有关他们年龄和性别的信息。他们可以精确地回答一个问题并

评估像性别和年龄这样的变量。种族划分的争论是一个比性别和年龄更复杂和抽象的变量。它典型地包含了物质、文化和历史因素，因此比性别或年龄要复杂得多，绝不仅仅是一个社会建构问题。虽然定义一个人的种族划分情况复杂又费时，但绝大多数人都能达到自我定义，能通过稍微的沉思或内省来报告他们的种族。因此，一个单一的变量在大多数情况下足以进行种族的划分。尽管如此，其他的许多理论变量都要求被试重建、解释、判断、比较或者评估不太容易得到的信息。例如，对已婚的人就以下问题进行测试，即如果他们选择不同的配偶的话其生活也会不同，很可能要求有大量的心理活动的参与，而且仅凭一个题项，可能也解决不了所关注的现象的复杂性。在这样的情况下，量表可能是最合适的测量工具。多题项可能会抓住此类变量的本质，并达到单一变量所不能达到的精确度。就是这种不能直接被观测到而且又需要被试思考的变量，最适合用量表进行评估。

量表应该与其他能产生合成分数的多题项测量类型相对照。这些不同的题项合成类型之间的差别同时具有理论的和实践的重要性，本书的后几章将会揭示这一点。在这本书所用的术语中，“量表”包含了伯伦(Bollen，1989，pp. 64～65；也可参见：Loehlin，1998，pp. 200～202)所称的“效果指标”(effect indicators)——即项目的价值由一个潜在的结构(或者我在下一章将会涉及的“潜在变量”)决定。抑郁量表通常符合量表的特点，即每个题项能估计行为的共同原因、行为名称和被试的情绪表达。因此，一个人怎样回答诸如“我感到很悲伤”和“我的生活很快乐”这一类问题，很可能主要取决于他当时的感觉。另一方面，我将用表征(index)这一术语来描述作为“原因指标”(cause indicators)的题项集，也就是能决定建构水平的那些题项。例如，对总统候选人吸引力的测量，可能符合这一类表征的特点。题项可能会从候选人的居住地、家庭规模、外表吸引度、激励竞选工作人员的能力以及潜在的经济来源做出评估。尽管这些特征可能没有共同的原因，但它们可能有共同的效果——增加总统竞选活动成功的可能性。单个题项不能决定测验的结果，但把他们联合起来就能达到目的。对能组合成一个合成分数的题项的分组来说，一个更一般的术语是突发变量(e-

mergent variable，例如，Cohen，Teresi，Marchi，Velez，1990)，它包括一些实体的集合，这些实体共享有某些特征并且在一个共同的类别标题下能够被归在一起。对特征进行分组未必就表明分在同一组内的特征之间就必然有因果联系。比如说，以少于 5 个字母的单词开头的句子很容易就被归在一起，尽管它既没有共同的原因也没有共同的结果。一个突发变量的“突然出现”仅仅是因为在研究的题项中发现某人或某事(正如一个数据分析程序)具有相似的类型。

不所有的量表都是等同的

遗憾的是，并非所有的题项合成都是认真地编制而成的。对许多量表来说，汇编(assembly)可能比编制(development)更合适。研究者常常匆匆拼凑或挖掘一些题项并假定他们能组合成一个合适的量表，并未考虑这些题项是否有共同的原因(因此产生一个量表)或有共同的结果(因此产生一个表征)。仅仅因为都是一个更上位水平类别内的成员，也并不意味着这些题项要么由共同的原因引起，要么导致共同的后果(因此建构一个突发变量)。

研究者可能在编制量表时不会利用理论，也可能因为错误地解释了一个量表所测量的内容而产生一个错误的理论结论。一个不幸的问题是，研究者在某一测量可能不能反映其所假定的变量时，便得出了某一建构是不重要的或某一理论内部是不一致的结论。之所以产生这种情况，是因为我们在研究中很少直接检查变量之间的关系。如先前所提到的，一个我们容易忘掉的事实是，许多有趣的变量并不能直接被观察到。可观察的中介(proxy)和不可观测到的变量之间会产生混淆。例如，乍想起来似乎可以直接观察到血压和体温两个变量，而我们实际观察到的是作为中介的水银柱。我们假定可观测的中介与它们打算表现的隐含变量之间是密切联系的。正如温度计这一例子，我们把温度计的水银水平作为“温度”，尽管严格地说，它只是对温度的一个可视化的表现(如热能)。实际温度和测量到的温度之间密切相关，而涉及的测量(水银所达到的测量价值)变量(热能量)几乎总是无足轻重的。当变量和它的指标之间的关系比温度计那个例子中的关系要弱得

多时，这种微弱的关系试图要揭示的现象可能使测量变得混淆起来，从而得出了错误的结论。考虑这样一个假设情形：研究者希望在现存的数据集上进行一个二次分析，我们假定研究者对社会支持在后来的职业成就上的作用感兴趣。研究者发现，可利用的数据集包含了许多有关被试在一段时间内的社会地位的信息，并且要求他们回答是否已经结过婚。事实上，在不同时间所收集的几个题项都是关于婚姻的。再做更深一步的假设，在缺乏能提供更详尽的社会评估数据的情况下，研究者会把收集到的婚姻题项组合成一个量表，并把它作为测量社会支持的量表。而许多社会科学家都认为，把社会支持和婚姻状况等同起来是不合适的。这一等同不仅会忽视社会支持的其他重要方面（如对受到的支持的性质的感知）也会包含潜在的不相关的因素（如测量时成人与儿童的地位问题）。如果研究者在运用这一评估方法的基础上，由假设得出这一结论——社会支持在职业成就中不起作用，那么就完全错了。事实上，这种对照是在职业成就和婚姻状况之间进行的，只有当婚姻状况实际显示了支持的水平，得出的结论才是有效的。

劣质测量的代价

如果一个最差劲的测量是惟一可以利用的测量，那么使用它的代价会比得到的好处要大得多。在社会科学中很少出现那种为了避免可怕的结果而立即采取措施的情境。当别无选择时，只能设法把手边上的最好的工具拿来应付。然而，甚至在这些很少见的情境中，在运用劣质量表进行测量时，其固有的难题并没有消失。使用不能评估假设所需要评估的内容的测量，会导致错误的结论。这是否意味着我们只能使用那些经历过严格的编制和具有广泛的效度的测量工具呢？未必。尽管在某些情境中，有缺陷的测量可能比没有测量要好得多，但我们应该认识到什么时候我们的测量程序是有缺陷的，并相应地调整我们的结论。

研究者常常会认为，相对于那些研究中的重要的科学问题而言，测量是不那么重要的，因此他们会努力通过较少使用量表而达到“有效利用”量表的目的。尽管如此，大量的测量仍然是有效研究的必要条件。研究者应该争取把他们感兴趣的理论建构和他们

所实施的测量方法等同起来。劣质的测量极大地限制了研究结论的有效性。对于一个较关注实际问题而对测量本身不感兴趣的研究者来说，从一开始就尽可能使研究的测量正确无误，这是十分重要的，并且在以后的研究中应该把它当作是理所当然的事。

研究者为了降低被试的负担也会错误地利用太过简单的量表。但事实上，即使是半数的被试完成的可靠量表，也要比全部的被试完成的不可靠的量表会产生更多的信息。如果你不能确定数据的含义，那么所收集的数据量便失去了意义。因此相对于被试完成的能产生有效数据的更长版本的量表，能够便捷地完成但不能产生有意义的信息的量表只不过是浪费时间和精力。

总结与展望

本章强调测量是包括行为和社会科学在内的所有的科学分支中的一个基本活动。心理测量学，作为社会科学中一个关注社会及心理现象的测量的特殊领域，其历史可追溯到远古时代。在社会科学中，理论在量表的编制过程中起着至关重要的作用，而在量表编制过程中，题项的分组反映了潜在的理论变量的水平。尽管如此，并非所有的题项集都能在这个意义上合成量表。编制量表比随意地挑选题项要复杂得多。利用不恰当的测量常常得不偿失。

后面的章节将更详尽地讨论量表编制的基本原理和方法。第2章探讨了“潜在变量”，即一个量表试图量化的潜在建构，它是后面章节所描述方法的理论基础；第3章为量表的信度以及隐藏的信度系数提供了一个概念化的基础；第4章则评价了量表的效度；第5章是关于量表编制的实际引导步骤；第6章介绍了量表编制的因素分析概念并描述了它们在量表编制中的作用；第7章对量表编制的方法之一——项目反应理论进行了一个概念性的介绍；最后一章则简单讨论了量表怎样适合更广阔的研究过程。

2

潜 在 变 量

Understanding the Latent Variable

结构与测量

作为题项值的假设性因素的潜在变量

路径图(path diagram)

测量模型的进一步阐述

平行测试(parallel tests)

其他的模型

练　习

本章主要通过一个概念图式(conceptual schema)来理解测量与这些测量所表示的结构之间的关系,虽然这并不是惟一可以利用的结构。项目反应理论则是我们将在第 7 章中要探讨的另外一种测量观点。因为其在概念方面和操作方面都具有相对的可利用性,并且使用广泛,因此我强调经典的测量模型,这种模型假设每一个题项都是潜在结构的可比较的指标。

结构与测量

一般来说,研究者对结构感兴趣,而不是对题项或量表本身感兴趣。例如,一个测量父母对孩子的期望的研究人员,会对父母的情感以及父母对孩子将来的成就所抱的希望这些无形的东西感兴趣,而不是对父母在问卷上所做的那些符号感兴趣。然而,在很多情况下,记录下被试对问卷的回答将会是测量这些情感与期望的最好方式。换句话说,这是必需的,因为很多结构都无法被直接测量。在某种意义上,测量是我们所无法直接观察到的很多变量的代表。通过评估各个测量之间的关系,我们可以直接推导出结构之间的关系。例如,在图 2.1 中,虽然我们的最初目的在于测量变量 A 和 B 之间的关系,但事实上我们是在与这两个变量有关的测量之间关系的基础上来评价二者之间的关系的。

一个量表要反映的潜在现象或结构通常被称为潜在变量(latent variable)。所谓潜在变量,顾名思义,它反映了两个主要的特征。我们以刚才所提到的父母对孩子的成就的期望为例来说明这两个特征。首先,它是潜在的,而不是显现的。父母对孩子的成就期望是不可以直接观察到的。此外,结构是可变的,而不是恒定的,它的某些方面,例如强度或大小(magnitude)在变化。父母对孩子的成就期望可能会随着时间(例如,在婴幼儿期间与青少年期间)、地点(例如,在运动场上与教室里面)、人物(例如,具有不同背景和职业的父母)以及其他因素的组合而变化。在这种对孩子的成就期望事例中,潜在变量是真正令人感兴趣的现象。虽然我们无法直接观察它或量化它,但是潜在变量在一些具体的环境条件

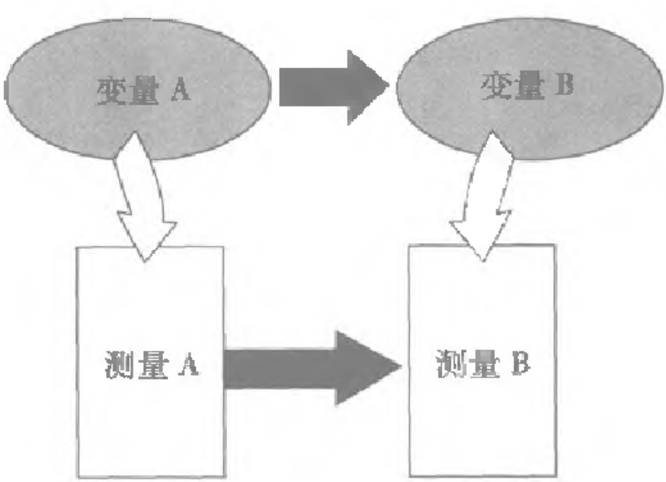


图 2.1 只有当每一个测量与其潜在变量相对应时，工具之间的关系才与潜在变量之间的关系相互对应

下，大概具有一个特定的值。所编制用来测量潜在变量的量表的目标在于，评估所测量的每一个被试在测量当时当地的实际大小。而这个无法观察的“实际大小”就是实际的分数。

作为题项值的假设性因素的潜在变量

潜在变量的概念表明了其与作用于它的题项之间的某种关系。潜在变量被认为是题项分数的原因，即是说，潜在变量的强度或数量(例如，它的实际分数的值)可能导致某个题项(或题项集合)具有某个值。

以下就是评估父母对孩子的成就期望的一些假设题目：

- 我的孩子的成就决定我自己的成功。
- 我愿意做几乎任何事情来确保我的孩子的成功。
- 如果有助于我的孩子取得成功，再大的牺牲都不为过。
- 对我而言，没有其他任何事情比我孩子取得成就更重要。

如果让父母表示在多大程度上赞同以上每一个题项的话，他

们对其孩子成就的潜在期望便会影响他们的反应。换句话说，每一个题项都应该给予潜在变量——对孩子的成就的期望——的强度一个指标。而在每个题项上所获得的分数是由特定的时间及特定的潜在变量的强度或数量决定的。

潜在变量和测量之间的因果关系表明了某些实验关系。例如，如果一个题项值是由一个潜在变量造成的，那么在这个值与这个潜在变量的实际分数之间就应该是相关的。因为我们不能直接获得这个真实分数，所以不能计算其与题项之间的相关。然而，当我们考察可能由同一潜在变量引起的一整套题项时，我们就能考察它们之间的相互关系。因此，如果我们有几个像以上测量父母对孩子的成就期望那样的题项的话，我们就能直接看出它们之间是如何相关的，并且把潜在变量当作这些题项之间相关的基础，以及使用这一信息来推测每个题项与潜在变量的相关情况。稍后，我将阐述所有这些如何能从题项之间的相关而得到。首先我要介绍一些图表方法，以使得这些阐释更加清晰。

路径图(path diagram)

在这里，与本问题有关的内容仅限于与量表的编制相关的话题。对于该问题的更深入探讨，请参阅阿希尔(Asher，1983)和洛林(Loehlin，1998)的研究。

图表惯例

路径图是一种用来描述变量之间因果关系的方法。虽然它们能够与在数据分析方法中的路径分析一起使用，但是，路径图作为详细说明一套变量是如何相关的方法有更广泛的用途。这些图表需要遵循某些惯例。从一个变量标签指向另外一个变量标签的一条直线箭头表明这两个变量是因果相关的，并且结果所在的方向就是箭头所指的方向。因此，X→Y 清楚地表明 X 是 Y 的原因。通常，联想路径(associational path)也由变量标签来确定，例如图 2.2 中的字母“a”。

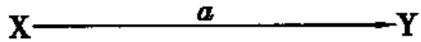


图 2.2　从 X 到 Y 的因果路径 a

而箭头的缺失也有其隐含意义，即这两个变量是不相关的。因此：A→B→C　D→E 表明 A 是 B 的原因，B 是 C 的原因，C 和 D 不相关，而 D 是 E 的原因。关于路径图的另外一个惯例是表示"误差"的方法，而该误差通常被描述为一个额外的原因变量。这个误差项(Error term)是"残差"(Residual)，它表示所有在图表中不能被所明确表述的原因解释的变化的原因。

因为误差项是残差，因此根据我们关于 X 和 Z 的知识(在该事例中)，它代表了 Y 的实际值和我们所预测的 Y 的值之间的偏差。有时候，这个误差项是假想的，因而并没有包括在图表中(图 2.3)。

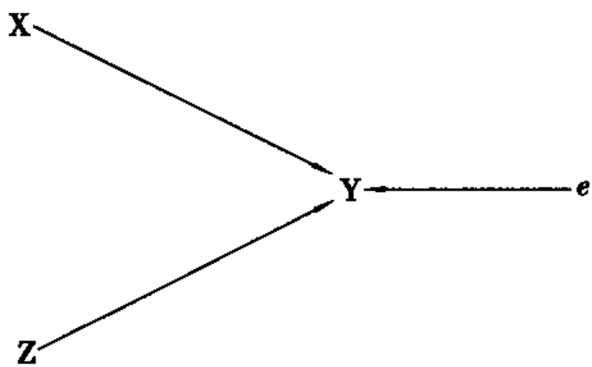


图 2.3　两个变量加上误差决定 Y

量表编制中的路径图

路径图能帮助我们清楚地看出题项是如何与潜在变量成因果相关的，也能够帮助我们理解题项间的某些关系是如何暗示了题项和潜在变量之间的关系的。我们从考察路径图的一个简单计算规则开始。让我们来看图 2.4 中的简单路径图。

路径中的数字是标准化路径系数(standardized path coefficient)。每一个数字代表由箭头所联结的变量之间的因果关系的强度。系数是标准的，这一事实意味着它们都使用相同的刻度来

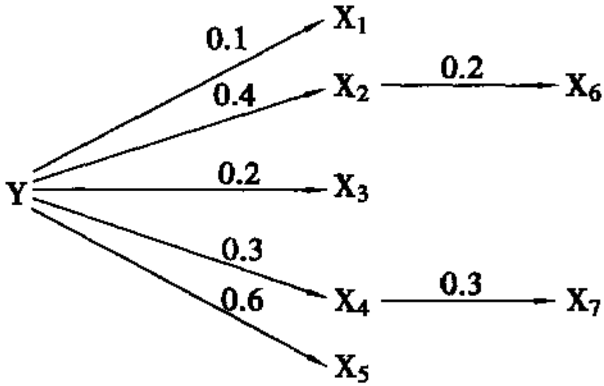


图 2.4　具有路径系数的路径图，可以用来计算变量之间的相关

量化因果关系。在这个图中，Y 是 X_1 到 X_5 的原因。在路径系数的值和 X_s（在量表编制型路径图中，代表题项）间的相关之间存在着一个非常有用的联系。对于像这个图表一样，只有一个共同源头（该例子中的 Y）的图表，任何两个 X_s 之间的相关值等于通过 Y 形成的 X 变量之间的路线中箭头上的系数的乘积。例如，X_1 和 X_5 之间的相关可以通过由 Y 把它们联结起来的标准路径系数的乘积来计算。因此，$r_{1,5}=0.6\times0.1=0.06$。变量 X_6 和 X_7 也共享一个 Y，尽管联结它们的路线要长一些。然而，规则仍然适用。由 X_7 开始，我们反向找到 Y，再往前寻到 X_6（或者，我们可以反过来，从 X_6 寻到 X_7），结果为：$0.3\times0.3\times0.4\times0.2=0.0072$，因此，$r_{6,7}=0.0072$。

路径系数和相关之间的这种关系为评估潜在变量和对之造成影响的题项之间的路径提供了一个基础。虽然潜在变量是假设性的并且不可测量，但是题项是实实在在的并且它们之间的关系可以直接计算。通过使用这种关系，即我们刚刚所讨论的简单规则以及关于这些题项与实际分数之间的关系的假设，我们可以得出关于题项和潜在变量之间的路径的估计值。我们可以从变量之间的一系列相关开始。然后，从路径与相关之间的关系往后计算，如果假设是正确的话，我们就能够计算出某个路径的值。让我们来看图 2.5 中的例子。

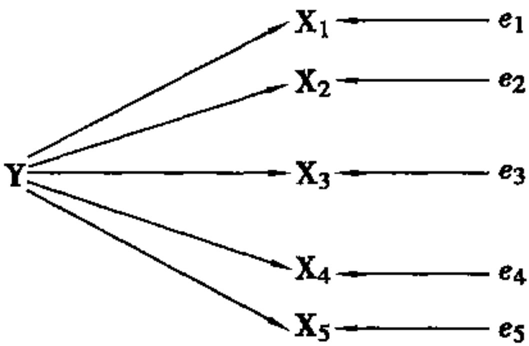


图 2.5　带有误差项的路径图

这幅图与先前所讨论的图相似，但在以下几个方面有所区别：没有路径值，变量 X_6 和 X_7 被去除了，剩下的 X 变量代表量表题项，并且每个题项都有一个变量（误差，用"e"来标注），而不是由 Y 来影响它。这些 e 变量在每个题项情境中都是独特的，并且表示不能由 Y 所解释的"残差"。这幅图表明，所有的题项都受 Y 的影响。此外，每个题项还要受一系列全部被当作误差的独特变量的影响。

这幅修订过的图表明 5 个单独的题项是如何与一个潜在变量 Y 相关的。e 和 X 的下标数字表明，这 5 个题项是不同的，并且与之一一对应的 5 个误差来源也是不一样的。在这幅图中没有箭头表示直接从一个 X 联结到另外一个 X，或者从一个 e 到另外一个 e，或者从 e 到与它没有联系的其他 X。它的这些特征是我们稍后将讨论的一些假设。

如果我们有一群人所完成的 5 个实际的题项，我们就会有这些题项的分数，并且我们可以得到它们之间的相关值。先前所讨论的规则使我们能够从路径系数来计算相关。加上其他的一些假设，它也可以使我们从相关来计算路径系数——即是说，从实际题项计算得来的相关能够用来决定每个题项是如何与潜在变量相关的。例如，如果 X_1 和 X_4 有一个相关值为 0.49 的话，那么我们就能知道从 Y 到 X_1 的路径值的乘积，并且从 Y 到 X_4 的路径值也等于 0.49。我们能够知道这些，是因为我们已经建立了的规则，即两个变量之间的相关等于联结它们的路径上的路径系数的乘积。如果

我们也假设这两个路径值是一样的，那么它们分别为 0.70。*

测量模型的进一步阐述

经典的测量假设

经典的测量模型首先有一些关于题项及其与潜在变量和误差来源的关系的假设：

- 与各个变量相联系的误差的数量随机变化。当与各个变量相联系的误差在大量的被试样本中合计时，其平均值为零。因此，当有大量的被试来完成题项时，题项的平均值几乎不受误差的影响。
- 一个题项的误差项并不与另外一个题项的误差项相关；联结题项的路径会经过潜在变量，但是决不会经过任何误差项。
- 误差项与潜在变量的实际分数不相关。注意，从潜在变量出发的路径并不向外延伸到误差项。题项与其误差项之间的箭头所指的方向相反。

以上前两个假设是作为很多分析程序的基础的一般统计假设，第三条实际上把误差定义为残差，即对预测值与结果，或者，题项与其潜在变量之间的所有关系充分考虑之后而余下的值。

平行测试(parallel tests)

经典的测量理论，传统上是建立在平行测试的假设之上的。平行测试这个术语来自于以下事实，即人们可以把每一个单独的

---

* 虽然 −0.70 也是 0.49 的一个平方根，但是在正根和负根之间作出一个选择，一般并不如我们所想的那样受关注。只要使所有的题项之间成正相关(如果需要，就如第 5 章所讨论的那样，通过对某些题项进行反向记分)，那么从潜在变量到每个题项之间的路径系数的符号就会一样，并且是任意的。然而要注意，给这些路径以正号暗示题项显示了结构以外的一些东西，而负的系数则有相反的暗示。

题项看作是潜在变量值的一个测试。就我们的目的而言，说“平行题项”可能更准确些。然而，从遵从习惯的角度讲，使用传统的名称更易为理解和接受。

平行测试模型的一个优点在于，根据我们对题项之间相关情况的观察，它的假设使我们非常容易得出关于每个题项与潜在变量是如何相关的结论。在此之前，我认为，如果有题项之间的相关方面的知识与一些假设，人们就能对从原因变量联结到题项之间的路径进行推测。就如在接下来的一章中所要讲到的那样，能够把一个数值分配给潜在变量和题项本身之间的关系是非常重要的。因此，在这一节中，我将详细地讨论平行测试的假设是如何得出这种可能性结果的。

作为平行测试模型基础的基本原理是，量表的每个题项既是对潜在变量的一个精确测量，也是对量表题项的任何其他成分的精确测量。因此，每个题项都是严格地平行的，即是说，每个题项与潜在变量的关系和每一个其他题项与该潜在变量的关系是完全相同的，并且出现在每个题项中的误差数量也是相同的。图 2.6 可以用来表示这个模型。

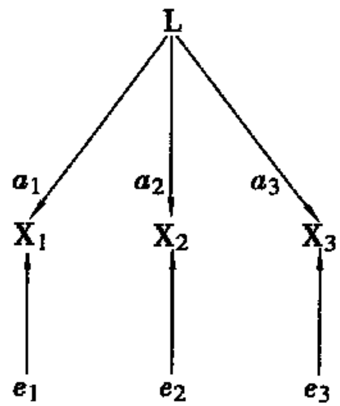


图 2.6　平行测试模型图，这里所有从潜在变量(L)到题项(X_1、X_2、X_3)的路径的值彼此相等，从误差项到题项也是如此

该模型在以前所列举的图表的基础上增加了两个假设：

• 潜在变量对每个题项的影响程度被假定都是一样的。

• 每个题项和其他的任何题项一样，都有相同的误差总量，即除了潜在变量以外的因素的影响在所有题项中都是相同的。

这些增加的假设意味着每个题项与分数之间的相关是相同的。能够断定这些相关相同这一点非常重要，因为这将有助于我们采用什么方法来决定这些相同相关的值。反过来，它也将决定量化信度的方法，这将在下一章中讨论。

断定实际分数与每个题项之间的相关是相同的，就需要前面所提到的两个增加的假设。平方相关(squared correlation)是两个变量之间所共有的方差的比例。

因此，如果实际分数与两个题项中的每一个之间的相关是相等的，那么该实际分数与每个题项之间所共有的方差的比例也是相等的。假设实际分数对两个题项的每一个所提供的方差数量是相同的。如果这些题项具有相同的总体方差，这个数量就可能是每个题项的总体方差的相同比例。为了使总体方差对两个题项来说都是相同的，每个题项从除了实际分数以外的其他因素中获得的方差数量也必须相等。由于除了实际分数以外的所有方差来源与之聚集在一起都称为误差，这就意味着这两个题项必须有相同的误差变化。例如，如果 X_1 从实际分数那里获得了 9 个任意的方差单位而从误差那里获得 1 个单位，那么实际分数的方差比例就是整个方差的 90%。如果 X_2 也从实际分数那里获得了 9 个方差单位，并且总体方差是 10 的话，那么这 9 个单位就是整个的 90%。如果像 X_1 一样，误差为 X_2 所提供的方差单位为 1，那么总体方差也只能等于 10。因而，每个题项与实际分数之间的相关就等于作用于实际分数的每个题项方差的比例的平方根，在这里大约为 0.95。

因此，由于平行测试模型假设从潜在变量而来的影响数量对于每个题项来说都是一样的，并且从其他地方(误差)而来的影响的数量也是相等的，所以对于所有题项来说，由于潜在变量和误差的作用所导致的题项方差的比例也是一样的。这也意味着，在平行测试的假设条件下，从潜在变量到每个题项的标准路径系数对于所有的题项来说也是相同的。标准的路径系数是相同的，这一假设使从题项之间的相关来计算路径系数成为可能，正如在前面的例子中所讨论的那样。前面所讨论的联结路径系数和相关之间

的路径图原则，会帮助我们理解当我们接受了先前的假设以后为什么这些等同性会存在。

这个模型的假设也表明，题项之间的相关是相同的（例如 X_1 和 X_2 之间的相关与 X_1 和 X_3，以及 X_2 和 X_3 之间的相关一样）。我们是如何从假设得出这样的结论的呢？之所以说相关全部等同，是因为解释任何两个题项之间的相关的机制是通过潜在变量而联结这些题项的路径。例如，X_1 和 X_2 仅仅是通过由 a_1 和 a_2 所组成的路径所联结的。二者之间的相关可以通过寻求联结该问题中的两个题项的路径并且把路径值相乘来计算。对于任何两个题项，这就是把有相同值（例如 $a_1=a_2=a_3$）的两个路径相乘。通过乘以相同的值而计算得来的相关当然会是相同的。

这个假设也表明，题项之间的每一个相关等于从潜在变量到一个题项之间的任何路径的平方。我们怎么得出这个结论的呢？两个不同路径（例如 a_1 和 a_2）的乘积等于每一个路径的平方，因为两个路径系数都是一样的。如果 $a_1=a_2=a_3$，并且 $(a_1\times a_2)=(a_1\times a_3)=(a_2\times a_3)$，那么每一个乘积一定等于自身相乘的任何一个 a-路径（a-paths）的值。

从这个模型的假设中我们也可以知道，与每一个题项相联系的误差的比例，是与潜在变量相联系的方差的比例的余数。换句话说，潜在变量所不能解释的对某个特定题项的任何影响一定能够由误差来解释。这两个影响一起对任何特定题项的方差做了100%的解释。这一点非常简单，因为误差项 e，被定义为包含除潜在变量以外的题项中的所有误差来源。

这些假设至少支持了另外一个结论：因为每个题项受潜在变量的影响是均等的，并且每一个误差项对相应题项的影响也一样，所以这些题项都有相同的平均数和方差值。如果能影响平均数的仅有两个来源对于所有题项都一样的话，那么很显然这些题项的平均数肯定也会是相等的。这个推理也适用于题项方差值。

总而言之，平行测试模型假设：

- 随机误差。
- 误差之间彼此不相关。
- 误差与实际分数不相关。
- 潜在变量对所有题项的影响相同。

• 每个题项的误差量相等。

这些假设使我们得出各种各样有趣的结论。此外，该模型使我们能够根据题项彼此之间的相关来推测潜在变量，但是，要达到这一点，该模型必须设定这四个严格的假设。

其他的模型

正如测量学所进展的那样，为了有效地推测实际分数和观察值之间的关系，与平行测试相伴随的所有严密的限制性假设是不必要的。一个建立在被称为“基本 Tau 相等测试”(essentially tau-equivalent test，有时候也叫做“随机平行测试”，randomly parallel test)基础之上的模型做出了较自由的假设，即是说，与特定题项相联系的误差变异的数量不必等于其他题项的误差变异(例如，Allen & Yen，1979)。因此，从潜在变量到每个题项的标准化的路径值可能不相等。然而，从潜在变量到每一个题项的非标准化的路径值(例如，与潜在变量对每个题项的影响的比例相对的数量)也被假设为对于每个题项都是相等的。这就意味着，在受潜在变量的影响但是不一定受完全同等程度的聚集在一起被称为误差的外在因素的影响程度方面，题项是平行的。在严格的平行假设下，不同的题项不但同等程度地作用于实际分数，而且它们的误差成分也是一样的。Tau 等价(在实际分数中，“tau”是希腊符号，等于“t”)更容易被承认，因为它不影响“相等误差”条件。因为误差可能会变化，因此题项值和方差值也有可能变化。对于该模型的更为自由的假设更有吸引力，因为要寻找到对于相同方差值的等同测量是很难的。这个模型使我们能够得出许多与我们用严格的平行测试但是限制性较少的假设所得出的结论相同的结论。读者可以把这个模型与农纳利和博恩斯腾(Bernstein)所讨论的“领域取样模型”(domain sampling model)相比较。

一些量表编制者认为，甚至基本的 Tau-等价模型也是限制性的。因为，我们不能经常假设每个题项受潜在变量的影响相同。研究者在所谓的同属模型(congeneric model，Joreskog，1971)指导下编制了一些测试来验证一套更为松散的假设(对于同属测试的

进一步讨论，请参考卡尔弥那斯和麦克艾维尔 1981 的研究，Carmines & McIver,1981)。它仅仅假设(超越了基本的测量假设)所有的题项都共享一个普遍的潜在变量。它们不必与潜在变量保持相同程度的关系，并且它们的误差变异也不必相等。但是必须假设每个题项在某种程度上反映了实际的分数。当然，每个题项与实际分数相关越紧密，量表就越可信。

一个不那么拘泥的方法是综合因素模型(general factor model)，该模型允许多个潜在变量作为一套特定题项的基础。卡尔弥那斯与麦克艾维尔(Carmines & McIver, 1981)、洛林(Loehlin, 1998)、隆(Long,1983)已经探讨了这种非常普遍化的模型的价值，主要是它与现实世界中的数据之间的一致性得到了改善。结构等式建模方式(structural equation modeling approach)常常把因素分析合并成其测量模型。多个潜在变量作为一套指标的基础这些情况为综合因素模型提供了例证(Loehlin, 1998)。

同属模型是因素模型的一种特殊情况(例如，单因素情况)。类似地，基本的 Tau-等价测量也是同属测量的一个特例——在这种情况下，题项与其潜在变量之间的关系被假设为相等的。最后，当增加一个假设，认为每个题项与其相关的误差源之间的关系相等时，严格的平行测试就是基本的 Tau-等价测试的一个特例。

此外还有一个测量策略应当被提到。这就是项目反应理论。在编制能力测试中，这种方法已经和二分反应(dichotomous-response，例如正确与不正确)题项一起被广泛使用，但它不是惟一的。在题项反应的更广类别中，不同的模型也许是以标准的，或者越来越频繁地，以逻辑的概率函数(logistic probability function)为基础。IRT 假设每个单独的题项对潜在变量有其特定的敏感性，用题项一特征曲线(item-characteristic curve, ICC)来表示。ICC 是潜在变量(如能力)的值与对一个题项的某个反应(如正确回答)的概率之间的关系的一个图示。因此，这条曲线反映了一个题项需要多少能力才能得以正确回答。我们将在第 7 章中进一步讨论 IRT。

除了在第 7 章要讨论 IRT 以及在第 6 章中要讨论因素分析以外，基于以下几个原因，我们将着重探讨平行和基本的 Tau-等价模型。首先，它们例证了经典的测量理论。此外，讨论其他模型得以

运行所依赖的机制会很快变得繁重。最后，对那些对测量有着初步兴趣的社会科学家来说，经典的模型是非常有用的，而对于那些对测量极为认真的人而言，却并非如此。这群人正是本书的读者。对于这些人，从一个经典的模型发展而来的量表编制程序会产生令人满意的量表。实际上，虽然就我所知道的而言，没有任何一个记分方式是现成可用的，但是我怀疑(能力测试以外)在社会科学研究中所使用的大量的众所周知的并被高度关注的量表是使用这些程序编制出来的。

练　习

1)根据两个题项之间的相关，我们怎么能够推断潜在变量和与潜在变量相关的两个题项之间的关系？

2)平行测试和基本的 Tau-等价模型之间在假设方面主要有什么不同？

3)哪个测量模型仅仅假设：在对所有的测量方法都普遍的基本假设之外，题项共享一个相同的潜在变量？

3 信　度

Reliability

连续题项与二分题项

内部一致性(internal consistency)

以量表分数间的相关为基础的信度

概括化理论(generalizability theory)

总　结

练　习

信度是心理学测量的一个基本话题。一旦它的含义得以充分理解，其重要性就显而易见了。量表的信度是指与潜在变量的实际分数的方差比例。虽然有很多计算信度的方式，但是它们都立足于这个基本的定义。然而，人们怎么样计算和操作信度会随着人们所使用的计算方法而变化。

连续题项与二分题项

虽然题项可能有各种各样的反应形式，但是在本章中我们假设题项反应由多值反应选项(multiple-value response options)所组成。二分题项(例如，只有两个反应选项的题项如"是"或"否"，或者有多个反应选项并且这些选项能够被分为"正确"与"错误")在能力测试中被广泛使用，个别情况下也在其他测量情境中使用。例如：

(1)苏黎世是瑞士的首都。　①正确　②错误

(2)P 的值是多少?　①1.41　②3.14　③2.78

利用二分反应的计算简单性来计算信度的很多方法已经编制出来了。一般的测量课本，如农纳利和博恩斯腾(Nunnally & Bernstein，1994)所编著的课本，都详细地介绍了这些方法。这些方法在测评信度的逻辑性方面，在很大程度上能与应用于多点、连续量表题项的更一般的方法相媲美。为了使行文简短些，本章将简要提及有关由二分题项所组成的量表的信度评估。这种量表的一些特征将在第 5 章中介绍。

内部一致性(internal consistency)

内部一致性信度，顾名思义，与一个量表中的题项的同质性有关。以经典的测量模型为基础的量表的目的在于测量一个单一的现象。正如我们在前面一章所见到的那样，测量理论表明，题项之

中的相关与题项和潜在变量的相关之间有某种逻辑联系。如果一个量表的题项与其潜在变量之间有很强的相关的话，那么它们彼此之间也有很强的相关。虽然我们无法直接观察到题项与潜在变量之间的相关，但是我们肯定能够确定题项之间是否彼此相关。一个量表的内部一致性程度会影响其题项之间相关。什么能够解释题项之间的相关呢？有两种可能：题项之间具有因果关系（例如题项 A 是题项 B 的原因）或者题项之间具有一个共同的原因。在大多数情况下，前者的解释是不可能的，从而使后者成为更明显的选择。因此，高的内部题项相关表明，这些题项都测量了同样的东西（例如，是其表现）。如果我们做出如前面一章中所做的假设，我们也能得出这样的结论：题项之间的高相关表明题项和潜在变量之间的高度相关。因此，一个复合量表的单维度量表应该由一套相互关联的题项所组成。测量多种现象的复合量表——例如，多维健康控制点量表（multidimensional health locus of control, MHLC, Wallston et al., 1978）——实际上就是相关量表的一类；每个"维度"就是一个单独的量表。

阿尔法系数(coefficient alpha)

内部一致性通常等价于克若恩巴齐（Cronbach，1951）的阿尔法系数，α。基于以下几个原因，我们将详细讨论一下阿尔法。首先，作为对信度的一个测量，它被广泛使用。其次，它与信度的定义之间的联系，与我们稍后所要讨论的其他情况下的信度测量（例如其他形式的方法）相比，更具有不证自明的特点。因此，对于那些对其信度的内部原理不熟悉的人来说，阿尔法可能会比其他信度计算方法更为神秘。最后，对阿尔法计算所隐含的逻辑的探索，为比较其他方法如何把握信度的本质提供了一个可靠的基础。

众所周知的库德尔—瑞查德松 20 公式（Kuder-Richardson 20 formula）就是二分题项的阿尔法的一个特殊版本（Nunnally & Bernstein, 1994）。然而，正如前面所述，我们将重点讨论应用于有多个反应选项的题项中的更普遍的形式。

你可以把一套题项分数中的所有可变性（variability）看作是由以下两者之一所导致的：①量表所测量的现象中的个体的实际变化（例如，在潜在变量中的实际变化）；②误差。事实就是如此，因

为经典的测量模型把"现象"(例如病人对于控制他们与医生之间互相影响的期望)看作是在量表分数中所有相同的变化的来源,而把"误差"看作是剩余的或不同的变化(例如无意造成的一个题项有两个含义)。考虑这一现象的另外一种方法是,把整个变化都看作有两个成分:信号(signal,例如,在病人对控制的期望中的真实差异)和噪音(noise,例如,除了由控制意愿所造成的真实差异之外的其他一切东西所导致的得分差异)。正如我们所要看到的那样,计算阿尔法把一套题项中的整个变化划分为信号和噪音两个成分。在整个变化中作为信号的比例等于阿尔法。因此,考虑阿尔法的另外一种方法是它等于 1 减去误差变化,或者,反过来,误差变化等于 1 减去阿尔法。

协方差矩阵(covariance matrix)

为了更充分地理解内部一致性,讨论一下一套量表题项的协方差矩阵是很有帮助的。一套量表的协方差矩阵反映了该量表作为一个整体的重要信息。

协方差矩阵是相关矩阵的一种更普遍的形式。在相关矩阵中,数据已经被标准化了,方差值被设定为 1.0;在协方差矩阵中,记录的数据没有被标准化。因此,在非标准化形式中,它包含了与相关矩阵相同的信息。协方差矩阵对角线上的因素是方差——题项自身之间的协方差——正如相关矩阵中的主对角线上的单位元素是变量的 1.0 标准化方差以及与其自身之间的相关一样。其对角线外的值是协方差,表达了标准化的变量组之间的关系,正如标准化中的相关系数一样。因此,从概念上来看,协方差矩阵由以下两方面组成:①单个变量的方差(在对角线上);②代表标准的变量组之间的非标准化关系的协方差(在对角线外)。

表 3.1 列举了 X_1、X_2、X_3 这三个变量的一个典型的协方差矩阵。

另外一个在某种程度上更简洁地使用惯例符号来表达矩阵、方差和协方差的方法如下:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 \end{bmatrix}$$

表 3.1　三个变量的方差和协方差

	X_1	X_2	X_3
X_1	Var_1	$Cov_{1,2}$	$Cov_{1,3}$
X_2	$Cov_{1,2}$	Var_2	$Cov_{2,3}$
X_3	$Cov_{1,3}$	$Cov_{2,3}$	Var_3

题项量表的协方差矩阵

让我们把注意力集中在一套组合起来就可以构成一个量表的题项的协方差矩阵的性质上。以上所呈现的协方差矩阵有三个变量：X_1、X_2、X_3。假设这些变量都是三个题项的实际分数，并且这三个题项，X_1、X_2、X_3，组合在一起时就构成了一个量表，我们称之为Y。这个矩阵能够告诉我们关于每个题项与作为一个整体的量表之间的关系的什么信息呢？

一个协方差矩阵有许多有趣的(至少是有用的)性质。其中之一就是，假设所有题项都具有相等的权重的话，那么把该协方差矩阵中的所有因素加在一起(例如，把对角线上的方差和对角线外的协方差加在一起)，就会给出一个数值，该数值刚好等于整个量表的方差值。因此，如果我们把符号化的协方差矩阵中的各项加起来的话，那么其结果数值就会是量表 Y 的方差值。这点非常重要，并且能够经受重复验证：如果假设所有题项权重相等的话，* 由所有题项所组成的一个量表 Y 的方差值等于该协方差矩阵中所有题项值之和。因此，由三个相同权重的题项，X_1、X_2和 X_3，组成的量表 Y 的方差值，与该协方差矩阵的题项之间有如下关系：$\sigma_y^2 = C$，即：

* 对于权重题项，协方差通过乘积而增加，方差通过它们相应的题项权重的平方而增加。关于这一点的更完整的叙述，参看农纳利的著作(1978，pp. 154～156)。

$$C=\begin{bmatrix}\sigma_1^2 & \sigma_{1,2} & \sigma_{1,3}\\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3}\\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2\end{bmatrix}$$

想了解关于这节所讨论的话题的更多信息的读者，可以参考农纳里(Nunnally，1978)对协方差矩阵的讨论以及纳蒙波第瑞(Namboodiri，1984)对统计学中的矩阵代数的介绍。对于单个题项的协方差矩阵有很多其他有用的知识没有在这里讨论。关于题项的协方差矩阵的应用在波恩斯特德(Bohrnstedt，1969)的研究中有讨论。

阿尔法与协方差矩阵

阿尔法被定义为，一个量表中由共同的因素所引起的总体方差的比例，大概是潜在变量的实际分数，而该潜在变量是所有题项的基础。因此，如果我们想计算阿尔法的话，有一个量表的总体方差值以及作为“共同”方差的比例的数值是很有帮助的。协方差矩阵正是我们想做到这一点所需要的。

回想一下我们在第 2 章中所用来描述题项与潜在变量是如何相关的图，如图 3.1。

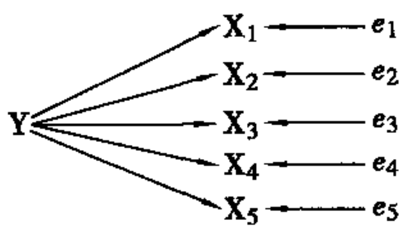


图 3.1　用图来表示有五个题项的一个集合如何与共同的潜在变量 Y 相关

题项中归因于潜在变量 Y 的所有方差都是被分享的或共有的(有时也用术语“共同的”或“公共的”来描述这种方差)。当 Y 变化时(例如，当其发生变化时，它表示各个题项会有不同水平的性质)，所有题项上的分数也会随之而变化，因为它是导致这些分数的原因。因此，如果 Y 值很高，那么所有这些题项分数也会很高；如果 Y 值较低，则它们也会较低。即是说，题项会共同变化(例如，

彼此相关)。因此,潜在变量影响所有题项,因而它们是相关的。相反,误差项则是每个题项所拥有的独特方差的来源。尽管所有题项都有由 Y 所引起的差异性,但是在我们的经典测量假设条件下,没有哪两个题项会都有来自于相同的误差源所造成的方差。某个给定误差项的值只能影响一个题项的分数。因此,误差项彼此之间并不相关。因而,每个题项(以及所暗含的由所有题项所共同组成的整个量表)作为以下因素的函数而变化:①其自身和其他题项所共有的方差来源;②我们称之为误差的、独特的、不共享的方差。由此得出结论,对于每一个题项以及因此而作为一个整体的量表的总体方差肯定是来自共同和独特因素的方差的一个复合。根据信度的定义,阿尔法应当等于共同因素方差与总体方差的比率。

现在,我们来考虑一下一个被称为 Y 的 k-题项(k-item)量表。其协方差矩阵如下:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots\cdots & \sigma_{1,k} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \cdots\cdots & \sigma_{2,k} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & \cdots\cdots & \sigma_{3,k} \\ \vdots & \vdots & \vdots & & \vdots \\ \sigma_{1,k} & \sigma_{2,k} & \sigma_{3,k} & \cdots\cdots & \sigma_k^2 \end{bmatrix}$$

该 k-题项量表的方差 σ_y^2,等于所有矩阵元素之和。主对角线上的记录是矩阵中所表示的单个题项的方差。第 i 个题项的方差用符号表示为 σ_i^2。因此,主对角线上的元素的和 $\sum\sigma_i^2$,是单个题项的方差之和。因此,协方差矩阵为我们计算以下两个值提供了现成的通路:①量表的总体方差 σ_y^2,被确定为矩阵中所有元素的和;②单个题项的方 差之和 $\sum\sigma_i^2$,通过把主对角线上的记录相加来计算。这两个值可以给予一个概念上的解释。从定义来看,整个矩阵之和即 Y 的方差,是由单个题项所组成的量表。然而,正如我们所说的,这个总体方差能够被划分为不同的部分。

通过探究在主对角线上的元素与所有对角线外的元素的不同,我们来探讨一下协方差矩阵是怎样区分共同方差与独特方差

的。所有的方差(对角线元素)都是单一变量或者"自身变量"(variable-with-itself)题项。我早先就已经说明,这些方差能够被看作是题项与其自身的协方差。每个方差仅仅包含一个题项的信息。换句话说,每一个方差所代表的信息都是以单一题项为基础的,而不是题项之间所分享的共同方差(在这个单一题项之内,其中一些变化会归因于共同的潜在变量,因而会与其他题项共享;一些则不会。然而,该题项的方差不会量化共享方差的程度,而仅仅是那个题项的分数中的离差量,也不考虑这种离散是什么造成的)。协方差矩阵中对角线以外的元素都涉及题项组,因而涉及两个量表题项之间共同的或联合的方差(协方差)。因此,协方差矩阵中的元素(因而 Y 的总体方差)由协方差(如果你愿意的话,也可以用共同方差)加上与单独考虑的题项有关的非共同的或非共有的方差所组成。图 3.2 表示了协方差矩阵的这两个细分部分。对角线中的共享区域是矩阵的非共有部分,而在对角线以外、三角边界内的两个区域一起,是共有部分。

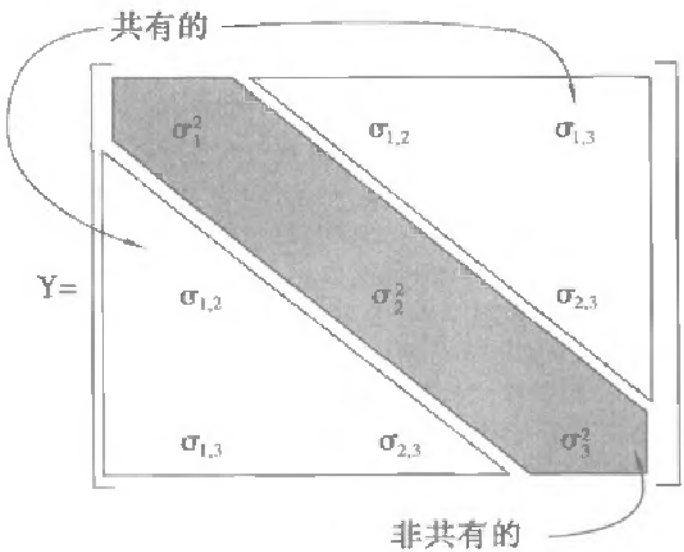


图 3.2 方差-协方差矩阵表明主对角线上的方差(阴影部分)是非共有的,而对角线之外的协方差(非阴影部分)则是共有的

由于有且只有协方差表示共有的变化，因此所有的非共有的变化一定要被表示在协方差矩阵的主对角线上的方差之中，因而由$\sum\sigma_i^2$来表示。总体方差，当然是由σ_y^2来表示，即所有矩阵元素的和。因此，我们把Y中的非共有方差与总体方差的比率表示为：

$$\sum\sigma_i^2 / \sigma_y^2$$

这个比率相当于协方差矩阵中的对角线值之和。由此可以得出结论，我们能够把共同的或共有的方差比例表示为其所剩余的部分，即是说这个值的补集，表示为：

$$1-\left(\sum\sigma_i^2 / \sigma_y^2\right)$$

这个值相当于协方差矩阵中所有对角线外的值之和。计算对角线元素并把它们从作为一个整体的协方差矩阵的值中减去，并不是最理想的方法。为什么不把对角线以外的元素的和直接计算为$\sum\sigma_{i,j}$呢？这里，i和j分别表示在一个特定的协方差中所涉及的两个题项。事实上，通过直接计算对角线以外的元素之和，我们会得出完全相同的结果。以上那个包含1的减数的公式是那个计算机还不发达的时代的遗产。计算Y的总体方差以及单个题项i，可能是用作其他目的的操作而完成的。即使没有必要为了其他目的而计算这些方差，考虑一下所涉及的计算量吧！对于一个有20个题项的量表来说，就在计算21个方差（每个题项一个，另外整个量表还有一个）或者计算190个协方差（例如，矩阵中对角线以外的380个元素各一个，这些在对角线以上和以下各不相同）或者总体方差之间进行选择。因此，把共有方差量化为从总体方差中去除掉非共有方差而剩下的部分比最初显然要合理些。其公式是：

$$1-\left(\sum\sigma_i^2 / \sigma_y^2\right)$$

或者：

$$\sum\sigma_i^2 / \sigma_y^2$$

这两个公式所表示的值，乍一看似乎抓住了阿尔法的定义，即量表中题项的共同因素造成了总体方差的共有部分，我们假设这反映了潜在变量的实际分数。然而，我们仍然还需要一个修正。如果我们有五个完全相关的题项，并且如果我们考虑将会发生的情况的话，这种修正将会更为重要，将会提高我们编制量表的信度。这种情况中的相关矩阵应该由一个所有值都等于1.0的5×5的矩阵组成。由此，前一个等式的分母的值应该等于25，而其分子的值却只能等于20，因而得出其信度为20/25，或者0.80而不是1.0。为什么会这样呢？协方差矩阵中的元素的总体数量是 k^2，矩阵中非共有的元素(例如，那些沿着主对角线的元素)的数量是 k，因而那些共有的元素(所有那些不在对角线上的元素)的数量为 k^2-k。因此，我们最后一个公式中的分数有一个以 k^2-k 的值为基础的分子以及一个以 k^2 的值为基础的分母。为了调整我们的计算以便这个比率表达的是相对大小而不是在分子和分母中所分别相加的项(term)的数量，我们乘以表示共有方差的比例的整个表达式的值，从而抵消在所有相加起来的项的数量中的方差。为了达到这一目的，我们乘以 $k^2/(k^2-k)$，即 $k/(k-1)$。这就把阿尔法的可能值限定在0.0到1.0这个范围之内。在刚刚讨论的五个题项的例子中，用0.80乘以5/4得到适当的1.0。读者可能想心算一下其他大小的矩阵。但是结果很明显，当题项都是完全相关的时候，$k/(k-1)$总是能得出一个阿尔法为1.0的乘数。因此，我们得出了关于协方差阿尔法的一般公式：

$$\alpha=\frac{k}{k-1}\left(1-\frac{\sum\sigma_i^2}{\sigma_{yi}^2}\right)$$

总之，一个量表的信度等于量表中总体方差在题项中的比例；这个总体方差是由潜在变量引起的，因而也是共有的。计算阿尔法的公式表达了这一点，它指定了题项集的特有的总体方差的比例，然后从1中减去这个比例从而决定共有部分的比例，然后再乘以一个修正因数来为先前计算做出贡献的元素的数量进行调整。

计算阿尔法的另外一个公式

计算阿尔法的另外一个一般公式是以相关为基础的，而不是协方差。实际上，其使用了$\bar{r}$，即平均题项间相关。这个公式是：

$$\alpha=\frac{k\bar{r}}{1+(k-1)\bar{r}}$$

从逻辑上看，它是由以协方差为基础的计算阿尔法的公式推导出来的。我们从概念术语的角度来考察一下协方差公式：

$$\alpha=\frac{k}{k-1}\left[1-\left(\frac{\text{题项方差之和}}{\text{方差与协方差之和}}\right)\right]$$

注意，右边项中的分子和分母是单个值之和。然而，这些单个值之和等于平均值乘以所涉及的值的数量（例如，k 个数字加起来等于 50，因此 k 乘以那些数字的平均值也等于 50。为了进一步证明这一点，用 10 来替代前一句中的 k；加起来等于 50 的 10 个数值的平均值等于 5，而 10 乘以 5 等于 50，与原始的和相等的数值）。因此，右边项的分子肯定等于 k 乘以平均题项方差$\bar{v}$，而分母一定等于 k 乘以平均方差加上(k^2-k)，或者 $k(k-1)$乘以平均协方差$(\bar{c})$：

$$\alpha=\frac{k}{k-1}\left[1-\frac{k\bar{v}}{k\bar{v}+k(k-1)\bar{c}}\right]$$

为了把“1”从这个等式中去掉，我们可以用与其等价的$[k\bar{v}+k(k-1)\bar{c}]/[k\bar{v}+k(k-1)\bar{c}]$来代替它，这就使我们把右边的整项化为一个比率：

$$r=\frac{k}{k-1}\left[\frac{k\bar{v}+k(k-1)\bar{c}-k\bar{v}}{k\bar{v}+k(k-1)\bar{c}}\right]$$

或者，与之等价的：

$$\alpha=\frac{k}{k-1}\left(\frac{k(k-1)\bar{c}}{k[\bar{v}+(k-1)\bar{c}]}\right)$$

从左边项的分子和右边项的分母中约掉 k，同时从右边的分子和左边的分母中一起约掉 $(k-1)$，得到一个简化的表达式：

$$\alpha=\frac{k\bar{c}}{\bar{v}+(k-1)\bar{c}}$$

我们所寻求的公式涉及相关而不是协方差，因而是标准化的而不是非标准化的项。标准化之后，协方差的平均值就等于相关的平均值，并且方差值为 1.0。因此，我们可以用题项间相关 $\bar{r}$，来代替 $\bar{c}$，用 1.0 来代替 $\bar{v}$。这就得出了以相关为基础的计算阿尔法系数的公式：

$$\alpha=\frac{k\bar{r}}{1+(k-1)\bar{r}}$$

这个公式被称为斯皮尔曼-布朗预占公式（Spearman-Brown prophecy formula），而其重要的用途之一将会在本章中讨论分半信度的计算时得以体现。

这两个不同的公式，一个以协方差为基础而另外一个以相关为基础，有时候分别指计算阿尔法的原始分数公式和标准化分数公式。原始分数公式在计算过程中保存了题项均值和方差值的信息，因为协方差是以保留了原始数据的原始测量的值为基础的。如果题项有显著不同的方差，当这个公式被用来计算阿尔法时，那些有较大方差的题项将比那些有较小方差的题项被给予更大权重。而以相关为基础的标准化的分数公式并不保留题项的原始测量公制（scaling metric）。相关是一个标准的协方差。因此，在标准化公式中，所有的题项被放在一个共同的公制上，因此在阿尔法的计算中权重相等。哪一个更好，这取决于特定情境以及是否需要相等的权重。为编制题项而建议使用的程序要求构建它们的措辞体系，以便使题项之间产生可比较的方差。这一点，我们将在接下

来的章节中将谈到。当按照合理的程序运算时，由这两种方法所计算出来的阿尔法系数中的方差一般很小。而当其目的在于产生相等的题项方差的程序未被遵循时，我们就会看到标准化的和原始的阿尔法值有明显方差(例如，0.05 或更多)，这预示着至少有一个题项有与另外一个题项的方差明显不同的方差。

信度与统计能力

与信度较低的量表相比，信度高的量表增加了特定大小样本的统计能力(或者使一个比之较小的样本产生了相等的统计能力)。例如，为了使两个实验组之间的一个特定数量的方差有一个特定的信度，我们需要一定大小的样本。得到这样一个方差(例如，测量的统计能力)的可能性可以通过增加样本大小而增大。在很多应用中，通过提高测量的信度，也可以达到同样的效果。一个有信度的测量，与较大的样本一样，会给统计分析造成较少的错误。在两者都可以运用的研究情境中，研究者应该努力衡量增加量表的信度与增加样本大小的相对优势。

通过提高信度而获得的统计能力依赖于许多因素，包括最初的样本大小、所设定来检测I类(Type I)错误的概率水平、所设定的显著效果的大小(例如，平均方差)，以及造成量表无信度的误差方差比例(而不是样本异质或其他因素)。要精确地比较信度增强与样本尺度增大之间的关系，就要求以上这些因素具体化。以下这些例子就证明了这一点。假设有这样一个研究情境，I类错误的概率水平被设定为 0.01，两种方法之间的 10 点(10-point)差异被认为是重要的，误差方差等于 100，而样本大小必须从 128 增加到 172(增长了 34%)，以便提高 F 检验的能力，从 0.80 提高到 0.90。减少整个误差方差，从 100 到 75(减少了 25%)，会产生完全同样的结果而没有增加样本的大小。用一个有较高信度的量表来代替一个信度较差的量表也可以达到这一效果。例如另外一个例子，$N=50$，信度为 0.38，并且其相关值($r=0.24$)，仅仅在 $p<0.10$时达到显著的两个量表，如果它们的信度增加到 0.90 的话，它们会在 $p<0.01$ 时达到显著。如果信度仍然保持在 0.38，那么要在$p<0.01$时达到显著的话，则需要一个两倍大的样本。利普塞(Lipsey，1990)提供了一个关于统计能力的更广泛的讨论，包括测量信度的作用。

以量表分数间的相关为基础的信度

除了内部一致性信度外，还有其他一些信度。计算信度的这些方法涉及在多种情况下，相同的被试完成一个量表的两个单独的版本或同一个版本。

信度的交替形式(alternative forms)

如果一个量表存在两种严格平行的形式的话，那么只要相同的人都完成了这两个平行的形式，就可以计算它们之间的相关了。例如，假设研究者最初编制了两套相同的题项，目的在于测量当病人与医生相互交流时其对控制的欲望，然后把这两套题项都运用到一组病人身上，最后求一套题项的分数与另外一套题项的分数之间的相关。这个相关就是交替形式的信度。这些平行形式是由题项所组成的，而所有这些题项(无论是在形式内部还是形式之间)都同样很好地测量了潜在变量。这就表明，这个量表的两种形式都有相同的阿尔法、平均值以及方差值，并且测量了相同的现象。本质上，平行形式由一套题项集合所组成，这些题项或多或少任意地被分成两个子集，而这两个子集构成了量表的两个平行而交替的形式。在这些条件下，一种形式与另外一种形式之间的相关值就等于每一种形式与自身之间的相关值，因为每一种交替形式就等于另外一种形式。

折半信度(split-half reliability)

交替形式的信度存在着一个问题，即我们通常找不到能够严格地遵循平行测试的假设的一个量表的两种版本。然而，能够找到把同样的逻辑应用到一个单一的题项集合中去的其他信度评估方法。因为交替形式本质上是由一个单一的题项库所组成的，这些题项被分成两组，由此我们可以：①选取组成一个单一量表的题项集合(例如，一个没有任何其他形式的量表)；②把这个题项集合分成两个子集；③求这两个子集的相关来评估信度。

这种类型的信度测量称为折半信度。折半信度实际上是计算

方法的一个种类而不是一个单一的类型，因为有大量的方法可以把这个量表分成两半。一种方法是把题项的第一半与第二半相比较。然而，这种"前半后半分割"可能是有问题的，因为除了潜在变量的值以外的因素(换句话说，即误差来源)会对每个子集有不同影响。例如，如果所研究的问题中构成量表的题项分散在一个较长的问卷中，当被试在完成量表的第二部分时就会更疲倦。于是这种疲倦就会在这两部分之间系统地变化并且由此会造成它们之间显得较不相似。然而，这种不相似性不会是题项本身的一个特征，正如在量表的题项顺序中它们的位置一样。另外一些会使先测试的部分与后测试的部分产生差别的因素有：练习效应使被试随着实验的进行而回答得越来越好，没有完成整套题项，甚至可能包括从前面到后面印刷质量方面的变化这些琐事也会造成影响。由于疲倦，这些因素会降低两个部分之间的相关，这是因为量表中题项所呈现的顺序而不是因为量表题项的质量所造成的。由于这些因素的结果，对题项之间的相关值进行测量，会由于与题项质量不直接相关的因素而变得复杂，从而导致一个错误的信度评估。

为了避免由于题项顺序所造成的缺陷，我们可以评估另外一种类型的折半信度，被称为奇-偶信度。在这种情况下，由奇数构成的题项子集与由偶数构成的题项子集进行比较。这就保证了这两个子集中的每一个题项都包含量表分段(例如，开始、中间和结尾)中的一个相同数目。假设题项顺序是不相关的(例如，与一般的成就测试的"由易到难"顺序相反)，这种方法避免了许多与前半部分和后半部分这种分半有关的问题。

理论上来讲，还有很多其他的方法也可以获得分半信度。作为以上所讨论的组建题项子集的方法的替代物，另外两个方法是平衡分半(balanced halves)和随机分半(random halves)。对于前一种情况，我们将识别一些重要的题项特征(例如以第一人称措辞，题项长度，或者问题中的某个特定类型的反应表示的是特征的出现或者缺失)。在构成这个量表的两个部分中，每一个部分都有相同的特征。因此，研究者应该以某种方式分配这些题项，从而使每个子集都有相同数量的以第一人称措辞的题项，相同数量的短题项等等。然而，当考虑复合题项的特征时，平衡了一半就很难平衡另外一半。例如，如果长的第一人称题项比短的第一人称题项

要多的话，就会出现这种情况。为后面的特征实现了平衡，这必然会造成前者的不平衡。此外，还很难决定题项的哪些特征应该被平衡。

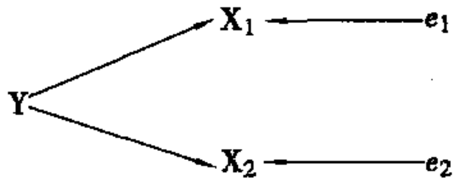


图 3.3 路径图表示一个测量的两个分割部分(X_1 和 X_2)与它们的共有潜在变量之间的关系

仅仅通过把题项分配给这两个子集中的任意一个，研究者就会取得随机分半，最后来求这两个子集之间的相关从而计算信度评估。这项工作成功与否，取决于题项的数量、所涉及的特征的数量以及特征中的独立性程度。期望一个小数量的题项，并且随着几个内部相关的维度变化，从而通过随机化得到可比较的分组，这是不现实的。但是，随机地把随着两个或三个不相关的特征变化的 50 个题项分配给两个类别，这样也会得到合理的可比较的子集。

采用哪一种方法来进行分半最好，这取决于特定的情境。最重要的是，研究者应当考虑，怎么样划分题项会导致不相等的子集，并且能够采取什么步骤来避免这种情况发生。分半信度和交替形式的信度，关于这两者背后的论证，都是平行测试模型的一个自然延伸。

当我们最初讨论一个模型时，虽然我们把每个题项当作一个"测试"，但是我们也可以把与这个模型一致的一个量表(或者一个量表的两个部分)看作是一个"测试"。因此，我们可以把在多个题项情况中所使用的原理运用到一个量表的两个交替形式或者两个分半情况中去。在平行测试假设下考虑两个"测试"(量表分半或者交替形式)。

从潜在变量到每一个变量之间的表示因果关系的路径，组成了连接这两个变量之间的惟一路径。因此，这些路径的值的乘积等于这两个测试之间的相关值。如果路径值必须相等(并且，在这

种模型的假设下，它们确实相等)，那么这两个测试之间的相关值就等于从潜在变量到任意一个测试之间的路径值的平方。这个路径的平方(假设它是一个标准化的路径系数)也是受潜在变量所影响的任意一个测试中的方差比例。反过来，这也就是信度的定义。因此，这两个测试之间的相关值就等于每一个测试的信度。

鉴于前一段中所指的"测试"是在交替形式情况中的一个量表的两个完全版本，因而它们在分半情况中则是两个分半量表。因此，求两个分半部分之间的相关，就得到了整套题项的每一部分的信度评估，但是这是对整个题项集合的信度的一个低估。以量表的一个部分的信度为基础，对整个量表的信度的评估，可以通过在本章中先前所讨论的斯皮尔曼-布朗公式来计算。回想一下这个公式：

$$\alpha=\frac{k\bar{r}}{1+(k-1)\bar{r}}$$

这里 k 是问题中的题项数量，而 $\bar{r}$ 是任何一个题项与另外一个题项之间的平均相关(例如，平均内部题项相关)。如果你已经知道了一个题项子集的信度(例如，通过分半的方法)，并且知道作为这个信度的基础的题项数量(例如，整个量表中题项数量的一半)，你就可以用这个公式来计算 $\bar{r}$。然后，你可以把这个 $\bar{r}$ 值和整个量表中的题项数量带回到公式里。以在量表的分半中所计算出来的一个信度值为基础，得出的结果就是整个量表的信度的估算结果。如果你在斯皮尔曼-布朗等式中使用一点代数学知识，从而把它变成以下这种形式的话，问题就简化了：

$$\bar{r}=\frac{r_{yy}}{[k-(k-1)r_{yy}]}$$

这里 r_{yy} 是问题中的题项集信度。例如，如果你知道两个含有 9 个题项的分半部分的折半信度等于 0.50，你就可以照这样来计算 $\bar{r}$：

$$\bar{r}=\frac{0.9}{[9-8\times0.9]}=0.5$$

然后你可以在斯皮尔曼-布朗公式中使用 r＝0.5，k＝18 来计算整个有 18 个题项的量表的信度了。因此，整个量表的信度估算为：

$$\frac{18\times 0.5}{1+(17\times 0.5)}$$

其结果等于 9/9.5 或者 0.947(注意增加了题项数量增大了信度。很快地看一下斯皮尔曼-布朗公式，我们就明显知道，如果所有其他条件都一样的话，一个较长的量表比一个较短的量表更可信)。

暂时的稳定性

另外一个计算信度的双分数(two-score)方法，涉及量表的暂时稳定性，或者从一种情境保持到另外一种情境分数如何保持不变。通常，测试-重测信度(test-retest reliability)是使用来评估这一指标的方法。在测试与医生交流的病人对控制的欲望时，假设研究者仅仅编制了一套题项而不是两套题项，分两个不同的时间给同一组病人测试这些题项，然后把第一阶段所得的分数与后一阶段测试的分数求相关。作为这种类型的信度测量的基础，其基本原理是，如果一个测量的确反映了一些有意义的结构，那么在不同的阶段里它所评估的结构应该有可比性。换句话说，潜在变量的实际分数应该对在两个(或更多)时期中所观察到的分数有可比较的影响，而误差成分在整个量表的测试中应该不会保持恒定。结果，对相同个体所实施的一个量表的两个测试中所取得的分数的相关，应该表示潜在变量对所观察到的分数的决定程度。这就与信度的定义相同，是由潜在变量的实际分数所造成的方差的比例。

但是，这个推理的问题在于，分数随着过去的时间所发生的变化可能(或者不可能)与测量程序的误差倾向有关。农纳利(Nunnally，1978)指出，即使当所感兴趣的结构已经发生了改变时，题项的特征可能使它们产生暂时稳定的反应。例如，如果某个焦虑测试既受到社会期望的影响又受到焦虑的影响，那么分数可能会保持稳定，而不管其在焦虑中的变化如何。分数中的稳定性，在不同时期的测试中的高相关中被反映出来，但不会是在所研究的现象

中的恒定性表现。然而，这个现象不可能会变化，而所变化的只是测量中的分数。即是说，这个量表是不可信的。或者，实际上当现象本身已经变化了并且测量也完全跟随其发生了变化时，分数中的变化可能是由于非可靠性导致的。问题是，造成一个变化的出现或者一个变化的缺失的原因，除了测量程序的可靠性与不可靠性以外，还有各种各样的因素。克里和麦克格瑞斯(Kelly & McGrath,1988)确定了当我们在不同的时间检测同一量表的两套分数时被混淆的四个因素，它们是：①所研究的结构中的真实变化(例如，在被试样本中的平均焦虑水平中的净增长)；②现象中的系统波动(例如，焦虑中的变化，在一些恒定的平均值周围，作为白天时间的一个函数)；③由被试或测量方法中的方差而不是由所研究的现象所引起的变化(例如，疲劳效应造成题项被误读)；④由于测量程序固有的不可靠性而造成的暂时的不稳定性。在这些因素当中，只有第四个才是非可靠性。这些研究者也注意到，虽然像多特征-多方法矩阵方法(multitrait-multimethod matrix approach，将在下一章中讨论)这些方法能够有所帮助，但是绝不可能完全澄清这些因素。

这并不是说证明暂时的稳定性不重要。在任何研究情境中，假设(或证明)按不同时间分开的测试具有高相关，这样的陈述都会受到批评。然而，我们在这些情境中所寻求的稳定性，既是测量的稳定性也是现象的稳定性。当我们自以为现象已经保持稳定的时候，测试-重测试相关仅仅告诉了我们关于测量的情况。这种自信不是经常有保证的。因此，测试-重测信度虽然重要，但是最好认为它反映了关于一个现象的本质和测量的信息，而不单单是后者。把随着时间变化的分数中的恒定性看作是暂时的稳定性更好些，因为它并不像测试-重测信度所表明的一样，测量中的误差是我们所观察到的任何非稳定性的来源。

概括化理论(generalizability theory)

到目前为止，我们对信度的讨论一直集中在把所观察到的方差划分为由潜在变量的实际分数所造成的部分和误差部分。这一

节简单地介绍一个更一般的框架来划分误差和非误差来源中的方差。

在我们把一个更好的划分误差方差的思想应用到测量中去之前,让我们来考虑一下一个更一般的研究范例,在这个例子中检验了大量的变化来源。假设一个研究者想测试一个培训项目的有效性,该培训项目的目的在于提高专业成就。再假设该研究者在一个大样本的大学教授以及一个对比样本的艺术家中进行这个培训项目。该研究者还选了一些教授和艺术家作为对比组,他们不参加培训项目但是要和培训项目的参加者一起接受相同的成就评估。当对这一研究进行总结时,研究者可能得出结论认为,关于这些成就的观察结果,反映了系统变化的三个可以确认的因素的作用:①参加者与非参加者;②教授与艺术家;③这些因素之间的相互作用。在这种情境中合理的分析策略是对所获得的分数进行方差分析(ANOVA),在分析的时候,把这些引起变化的因素都分别当作一个维度。本质上,这种分析策略会把所观察到的关于成就的分数中的整体方差划分为几个来源:培训的参与,职业,它们之间的相互作用,以及误差。误差表示除了由前面的因素所确定的来源以外的所有差异来源。

现在,考虑一个假设的情境。在这个情境中,研究者正编制一个关于自治愿望的量表。量表的研究对象是年长的个体,并且这些人中的一些可能有视觉问题。接下来,研究者决定对那些阅读有困难的人进行口头上的自治愿望测验,而对其余的参加者则进行书面形式的测试。

此外,对于研究者来说,通过使用方差分析的方法,有可能承认作为分数中的变化来源的测试模式。如果分析结果证明了测试方法之间的差异解释了分数中的总体变化的不重要部分,那么在完成口语版或书面版的被试的分数的可比性方面,研究者就有更高的自信。另外一方面,如果分数中所观察到的总体变化的显著数量是由于测试模式造成的,那么研究者就知道,对于分数的任何解释都要考虑测试模式之间的这种差异。

概括化理论(例如,Cronbach, Gleser, Nanda, & Rajarat-

nam, 1972)提供了一个框架来检验我们所能假设的一个或多个维度的测试过程的相等程度。在前面的例子中,问题中的维度是测试模型。问题中的每个维度都是变化的一个潜在来源,并且被当作一个方面来考虑。这个例子集中在把模型当作变化的惟一潜在来源(除了个体)。在这个例子中,研究者希望概括化。因此,这个例子只涉及一个方面。

根据概括化理论的说法,在一方面的所有水平中(例如,这个量表既有口头形式的测试也有书面形式的测试)所获得的观察数据组成了一个可以接受的观察总体。这些观察的均值被认为是总体分数(universe score)并且类似于经典测试理论中的实际分数(Allen & Yen,1979)。概括化研究(generalizability study),或者G-研究,其目的在于决定一个方面的不同水平中的分数的可比较性程度。关于自治愿望这一假设性研究就是G-研究的一个例子,因为它表明了测试模型这一方面的不同"水平"的作用。

G-研究的目的在于帮助研究者测定某个方面对概括化的限制程度或不限制程度。如果一个方面(例如,测试模型)解释了在所观察到的分数中的方差的一个显著数量,那么其结果就无法概括那个方面的各个水平(例如,口头与书面测试)。在表示数据没有错误的情况下,我们可以概括该方面的各个水平的程度被表示为概括化系数(generalizability coefficient)。它通常是通过从合适的均值平方中形成一个比例来计算的,这个均值平方则来自于作为G-研究的一部分的 ANOVA。从概念上讲,概括化系数是总体分数方差与所观察到的分数方差之间的一个比率,并且与信度系数相类似(Allen & Yen,1979)。然而需要注意的是,如果一个G-研究得出了一个差的概括化系数,该研究的设计指向了引起该问题的一个原因,即所检测的那个方面。信度系数仅仅确定误差的数量,而并没有把之归结为任何一个特定因素。

在某些情况中,选择合适的 ANOVA 设计,决定哪些效果与问题中的方面相对应以及构建正确的概括化系数都是比较有难度的。正如一般的变量分析一样,复合维度、嵌套的、交叉的,以及混合的效应会使 G-研究复杂化(对于 ANOVA 设计的一般讨论,参

见 Myers,1979;或者 Kirk, 1995)。建议使G-研究的设计简单些。对于一个特定类型的G-研究,在考虑详细解释如何构建合适的ANVOA模型的原始资料的时候,也要谨慎。克若克尔和阿尔吉纳(Crocker & Algina,1986)描述了几个不同的单方面和两方面(one-and two-facet)的概括化研究的合适设计。这些资料也为概括化理论提供了一个好的总体介绍。

总 结

量表的可靠性与它们所包括的题项的可靠性程度一致,其题项都有一个共同的潜在变量。阿尔法系数与信度的经典定义对应紧密,认为信度是一个量表中由潜在变量的实际分数所引起的方差的比例。计算信度的各种方法在特定情境中有不同的效用。例如,如果我们无法得到一个量表的平行版本,计算交替形式的信度就不可能。在编制一个测量研究或评估一个已经出版的报告的时候,那些理解了计算信度的不同方法的优点与缺点的研究者,就能更好地得出见多识广的结论。

练 习*

1)如果一套题项有比较好的内部一致性,这暗示了有关题项与它们的潜在变量之间的关系的什么信息?

2)在这个练习中,**假设以下是一个量表的协方差矩阵,Y,由三个题项组成,X_1、X_2、X_3:

* 在整本书中,对于要求一个数字答案的任何练习的解答,都会在相应章的注释部分中找到,而练习也出现在相应章中。

** 答案是:a. 1.2, 1.0, 1.8(加起来等于 4.0);b. 7.0(矩阵中所有元素的和);c. (3/2)×[1−(4.0/7.0)]=0.64。

1.2	0.5	0.4
0.5	1.0	0.6
0.4	0.6	1.8

a. X_1、X_2和 X_3的方差是多少?

b. Y 的方差是多少?

c. 量表 Y 的阿尔法系数是多少?

3)用实际的量表特征,讨论一下测试—重测信度与其他因素之间在哪些方面有混淆。

4)交替形式的信度原理是如何遵循平行测试的假设的?

4

效　度
Validity

内容效度
标准-相关效度
结构效度
表面效度(face validity)
练　习

信度涉及一个变量对一套题项的影响程度，而效度涉及该变量是否是题项共变的潜在原因。一旦一个量表具有了信度，那么量表分数中的方差就可以归因于某些现象的实际分数，该现象对所有题项都有一个因果关系的影响。然而，确定一个量表具有信度并不能保证量表编制者感兴趣的变量实际上就是所有题项所共享的潜在变量。一个量表中，对一个特定变量（例如，知觉到的心理压力）的测量的充分性问题就是效度问题。

很多研究者对效度进行了不同的阐释。例如，麦斯克（Messick，1995）描述了六种类型的效度，其中之一（结果效度）涉及被试的分数如何对被试产生影响。虽然麦斯克（Messick，1995）对效度的评论产生了一些发人深省的问题，但是他的分类系统并没有被广泛采用。根据更为传统的解释，效度是从一个量表得以构建的方式、预测特定事件的能力，或者与其他结构的测量之间的关系之中推断出来的。与这些解释相对应，基本上可以把效度分为三种类型：

1）内容效度（content validity）。

2）标准-相关效度（criterion-related validity）。

3）结构效度（construct validity）。

在本章中，我将对每一种类型都进行简单地评述。对于效度的更为广泛的讨论，包括在标准相关效度以及其他效度指标中的方法学和统计学问题的讨论，参见盖塞利、坎贝尔和泽德克的论述（Ghiselli，Campbell & Zedeck，1981，第 10 章）。如果读者想了解关于效度的更为全面的观点，请参看麦斯克（Messick，1995）的著作。

内容效度

内容效度涉及题项取样的充分性问题——即是说，一个特定的题项集合（item set）对一个内容范畴（content domain）的反映程度。当这个范畴（例如，教给六年级学生的所有词语）得以很好地定义时，内容效度是最容易评价的。当测量诸如信念、态度或性格

这样的特性的时候，以及当一个样本题项具有代表性的时候，问题就会难解一些，因为很难确切地确定这些潜在题项的范围是什么。从理论上讲，当其题项是从一个大量合适的题项集合中随机选出来的一个子集的时候，一个量表就有内容效度。在以上所列举的词汇测试这个例子中，这一点非常容易实现。所有在学校期间所教的词语就会被定义为题项集合。于是就可以抽样出一些子集。然而，在诸如对信念的测量这种情况中，我们没有一个方便而合适的题项集合。尽管如此，我们编制量表的方法(就如第 5 章中所提供的建议那样，让专家就题项与研究的范畴之间的关联性对题项进行评价)能够有助于使题项的合适性最大化。例如，如果研究者需要编制一个量表来测量预期的结果与想要的结果(例如，在决策时，预期医生的决策与期望的结果)之间的差异的话，那么对他或她来说，使所有相关的结果都出现在题项中或许是令人满意的。为了做到这一点，研究者或许会邀请熟悉这个研究领域的专家来评价最初的题项清单，并请他们就那些已经被删除的但实际应该被包括进来的内容范畴提出意见。那么能够反映这个内容的题项就应该被添加进来。

标准-相关效度

为了获得标准-相关效度，顾名思义，就要求一个题项或量表与某个标准或“金标准”只有一个理论上的联系。这个联系的理论标准是否被理解与标准-相关效度无关。例如，如果我们能够表明探寻水源在理论上与确定地下水源的位置有关，那么探寻水源在关于成功地钻井挖掘这个标准上就有效度。因此，标准-相关效度本质上是一个实际的问题而不是一个理论的问题，因为其与理解过程无关，而仅仅与对它的预测有关。实际上，标准-相关效度经常被称为预测效度。

任何名义的标准-相关效度并不必然暗示着变量之间有一个因果关系，即使当对预测的情况和标准进行的排序是明确的时候，也是如此。当然，在理论范畴内的预测(例如，预测作为一个假设)，

可能与变量中因果关系有关并且可以为一个有用的科学目的所利用。

关于标准-相关效度，值得注意的另外一点是，逻辑上我们正在处理多种类型的效度问题，即标准是追随还是领先，或者是同步于问题中的测量？因此，除了“预测性效度”以外，并发性效度（concurrent validity；例如，在驾驶测试的同时通过口头询问问题并要求被试回答，从而“预测”其驾驶技能），或者甚至后预测效度（postdictive validity；例如，从婴儿发展状态量表中“预测”其出生时的体重）或许会或多或少地与标准-相关效度同时使用。标准-相关效度的最重要的方面，并不是问题中的测量与我们试图推测其分值的标准之间的时间关系，而是这两个事件之间的实验上的相关强度。标准-相关效度这个术语比其他暂时不确定的术语要优越，因而也更可取。

标准-相关效度与精确性

在结束标准-相关效度的讨论之前，有必要讲一讲其与精确性之间的关系。正如盖塞利（Ghiselli, 1981）和其同事所指出的那样，当考虑预测的精确性这个问题时，相关系数一直以来就是标准-相关效度的传统指标，但可能不是很有用。例如，一个相关系数并没有揭露出有多少情况是由一个预测指标所正确地分类的（虽然盖塞利等描述了一些表格，根据预测指标和标准之间的相关大小，这些表格提供了一个对这些情况落入不同百分比的类别的比例的估计，1981，p. 311）。在某些情况下，把预测指标与其标准都划分为离散的类别，并且对根据它们的预测指标把各种情况划分到正确的标准类别之中的“命中率”进行评价，这样可能更为合适。例如，我们可以把每一个变量划分到“低”与“高”的类别里去，并且把精确性概念化为正确分类的比例（例如，当预测指标的分值与标准的分值相对应时的情况）。一个需要重点考虑的问题是，我们在哪里分类。考虑一下以下情况：假如我们有两个固定状态的标准，例如“生病”与“健康”，以及有一个评价工具使研究者对分数进行二分处理。评价工具的目的是预测人们对于问题中的生病的反应是肯定的还是否定的。因为这个结果是二分的，所以使预测者也是

二分的这是有道理的。在分类中有两个可能的错误:量表可能错误地把一个实际上生病的人划分为一个健康的人(错误否定),或者把一个实际上健康的人划分为生病的人(错误肯定)。当二分会影响这两种错误类型的比例时,沿着评价工具的分数范围画一条分界线,在两个端点处,把任何人划分为健康的都会避免错误否定(但是会增加错误肯定),而把任何一个人划分为生病的都会避免任何错误肯定(但是会增加错误否定)。显然,在这两种极端的情况中,评价工具都根本没有任何预测价值。当然,其目标在于寻求一个能最少地产生这两种错误中的任何一种的分界,并且因此有最高的精确性。然而,常常是没有一个理想的分界点,即是说,难以找到一个能够完美分类的点。在这种情况下,研究者会有意识地最小化其中一个错误。例如,如果生病的状态是非常好的,治疗也是有效的,费用也便宜,并且病情是良性的,那么错误否定(导致治疗不足)的代价远远地大于错误肯定(导致治疗过度)的代价。因此,寻找到一个分界点以便减少错误否定而接受错误肯定似乎是合适的。另外一方面,如果这个治疗既昂贵又不舒服并且病情也比较严重,那么相反的选择或许更有意义。

同样,即使一个预测量表与一个标准之间的相关是理想的,预测量表中所取得的分数也不一定是对这个标准的一个估计,记住这一点很重要。相关系数对一个变量或两个变量的线形转换是不敏感的。两个变量之间的高相关暗示着同一个体在这些变量上获得的分数在它们各自的分配中会占据相似的位置。例如,如果两个题项之间高度相关的话,那么在第一个题项中打分非常高的人也可能在第二个题项中打非常高的分。然而,像“非常高”这样的题项却是一个相对的而不是绝对的题项,并且没有考虑对这两个变量的测量的统一部分。把测量的预测部分转换为标准部分对于取得一个精确的数字预测来说可能是必要的。这个调节等于剔除了一个回归线的倾斜之后再决定合适的截取。无法认识到一个分数需要转换,这可能会导致错误的结论。如果预测量表中与标准相同的部分碰巧是被校准过的,那么这种类型的错误或许很可能发生。例如,假设有人编制了以下“驾驶罚单量表”以预测驾驶员在5年内可能会收到多少张罚单:

(1)当我驾驶的时候我超过了限制速度。

经常　　　偶尔　　　少有　　　从不

(2)在多通道公路上,我驾驶在超车道中。

经常　　　偶尔　　　少有　　　从不

(3)我自己判断什么样的驾驶速度是合适的。

经常　　　偶尔　　　少有　　　从不

让我们也做一个不切实际的假设,即量表与在 5 年内所获得罚单的数量完全相关。这个量表的评分标准为:当一个被试选"经常"这个选项时这个题项的值为 3,"偶尔"的值为 2,"少有"的值为 1,而"从不"的值为 0。然后题项的分数被加起来得到一个量表分数。这个分数的理想的标准-相关效度并不意味着在 5 年内 9 这个分数转变为了 9 张罚单。相反,这意味着在这个测量工具上获得最高分数的人也是那些在一年之中被观察到的罚单数量最多的人。根据有些理论所做的转换(例如,0.33×分数)会得到实际的估计。这种特殊的转换将会为分数是 9 的驾驶员预测 3 张罚单。如果标准-相关效度很高,那么可能要计算一个更精确的估价才行。然而,在一个合适的转换之前的标准数值与预测数值之间的相似性可能与效度的水平毫无关系。

结构效度

结构效度(Cronbach & Meehl, 1955)涉及一个变量(例如,某个量表中的一个分数)与其他变量之间的实验关系。它表现了其所意欲测量的结构与已经建立的其他结构之间的相关程度。因此,如果我们根据理论把一些变量看作是与结构 A 和 B 正相关,与 C 和 D 负相关,并且与 X 和 Y 不相关,那么目的在于测量那个结构的量表应该对这些结构的测量有一个类似的关系。换句话说,我们的测量应该与 A 和 B 结构的测量正相关,与 C 和 D 的测量负相关,并且与 X 和 Y 的测量不相关。对于这些假设关系的描述可以参看图 4.1。

	A	B	C	D	X	Y
变量	+	+	−	−	0	0

图 4.1 变量间的假设关系

实验上的相关与预测的模式之间的匹配程度，为该测量对其意欲测量的变量的反映程度提供了一些证据。

结构效度与标准-相关效度的区别

人们经常混淆结构效度和标准-相关效度，因为同样精确的相关都能达到其中任何一个效度。它们之间的区别更多地在于研究者的目的而不在于所取得的分值。例如，流行病学家可能试图确定在调查研究中所取得的大量测量结果中的哪些因素与健康状态有关。其目的可能仅仅在于识别危险因素，而不考虑(至少最初)连接分数和健康状态的潜在的因果机制。这个例子中的效度就是这些测量能够预测健康状态的程度。此外，关系应该是更为理论性的和阐释性的。研究者，例如在本书的第 1 章中所描述的流行病学家，可能会认可把压力看作是影响健康状态的一个原因的理论模型，因而问题可能就是一个新编制的量表如何测量压力。这可以通过理论所认可的方式来评估，即理论上应该如何操作相关的量表“行为”来评价压力。如果理论认为压力和健康状态是相关的，那么在前一个例子中，被用作预测效度的证据的同一实验联系也应该被当作结构效度的证据。

根据研究者的意图，所谓的知名专家组的确认，也是一种能够区分结构效度或标准-相关效度的方式。知名专家组的确认典型地表现为，根据量表分数，证明某个测量能够把一个组的成员与另外一个组的成员区别开来。这个目的要么是与实验性有关的(例如，当能够正确地区别那些属于或不属于某个组的那些成员时，对这个组的态度的测量就是有效的)，要么是纯预测性的(例如，当某个人用一系列看起来不相关的题项来预测工作收入时)。在前一种情况中，这应该被看作是一种结构效度，而后者应该看作是一种标准—相关效度。

为了证明结构效度，相关应该为多高

我们无法找到分界点来确定结构效度。认识到两个测量可能不仅仅共享有结构相似性，这是很重要的。尤其是，在测量结构的方式中的相似性可能对分数中的共变（即结构相似性的自变量）进行解释的时候。例如，以多点记分系统（例如分值从 1 到 100）来记分的两个变量与一个双极变量相比，如果所有其他情况都一样的话，它们之间会有更高的相关。这是由测量方法的结构相似性所造成的人为产物。同样，由于程序的相似性，由测试者所收集的一种类型的数据与以同样的方式所收集的其他数据有某种程度的相关。即是说，两个变量之间的某些共变可能是由于测量相似性而不是结构相似性造成的。这一事实有助于回答以下问题，即有没有必要考虑结构效度的相关程度。变量至少应该证明除了共同的方法造成的变化以外的共变。

多特征-多方法矩阵(multitrait-multimethod matrix)

坎贝尔和费斯克(Campbell & Fiske，1959)编制了一个叫作多特征-多方法矩阵的程序，这对于测量结构效度非常有用。这个程序用一种以上的方法来测量一种以上的结构，所以我们可以取得一个"完全交叉"的方法来构建测量的矩阵。例如，假设设计了一个研究，在这个研究中每一次使用两个不同的测量程序分别对焦虑、抑郁和鞋子的尺码进行两次测量（注意，如果同时测量两个不同的被试样本，这会对这个方法的基本原理产生什么影响）。每一个结构的测量都应该包括两种方法、一个形象化类比量表（一根线条，被试在其上面做个记号以表示他们所拥有的特征的数量，这些特征包括焦虑、压抑或者鞋子的尺码）以及测试者在与每个被试 15 分钟的交互活动之后的评分。然后，我们就可以构建一个在这些测量中所取得的相关的矩阵，如表 4.1。

表 4.1　对多特征-多方法矩阵中相关的解释

项目 2 \ 项目 1	A_v	A_i	D_v	D_i	S_v	S_i
A_v	TM	T	M		M	
A_i	T	TM		M		M
D_v	M		TM	T	M	
D_i		M	T	TM		M
S_v	M		M		TM	T
S_i		M		M	T	TM

注：TM 代表同样的特征和方法(信度)；T 代表同样的特征，不同的方法；M 代表同样的方法，不同的特征；A、D 和 S 分别指焦虑、抑郁和鞋子的尺码；下标 v 和 i 指形象化类别和询问方式。

在表中并没有表现出相关特征与无关特征之间的区别。因为反映同一特征(结构)和同一方法的题项应该既共享方法上的变化，也共享结构上的变化，所以我们可能会假设这些题项的相关程度很高。同一特征但是不同方法的相关程度其次。如果是这样，这就表明结构共变比方法共变的相关要高；换句话说，我们的测量更多地是受我们所要测量的内容的影响而不是受测量方法的影响。相反，当用不同的程序来测量它们的时候，鞋子的尺码与这两个结构中的任何一个没有任何理由会存在相关。因此，这些相关不会与零有显著方差。对于不确定的但在理论上有联系的结构，例如压抑和焦虑，我们会假设某种结构共变。它们之间的相关会潜在地为建立结构效度提供很多信息。例如，如果我们的抑郁量表都是构建的比较好的了，而我们的焦虑量表目前正在编制，我们就可以评价在相似的和不同的测量程序情况下由概念相似性所引起的共变的总量。从理论上来讲，即使当通过不同的方法来测量的时候，焦虑和抑郁都应该是充分相关的。如果这被证明为是事实的话，它就可以被用作我们新编制的焦虑量表的结构效度的证明。更特别地，这些相关应该是聚焦效度(convergent validity)的表现，即在理论上与结构相关的测量之间的相似性的证明。理想

地焦虑与抑郁之间的相关应该小于两个抑郁测量之间或者两个焦虑测量之间的相关,但是它们都会比抑郁测量和鞋子尺码测量之间的相关要大。不管测量方法的相似性或非相似性,焦虑测量与鞋子尺码的测量之间无显著的相关,这一证明也同样重要。因为这就是差别效度(discriminant validity,有时也叫作发散效度,divergent validity)的证明,即没有联系的结构的测量之间没有相关。当以同样的方式来测量时,如果鞋子尺码与焦虑之间有显著相关,这表明方法本身解释了大量的与不同结构的相同测量有关的方差量(以及共变量)。

麦特克尔(Mitchell,1979)观察到,在为多特征-多方法矩阵收集数据中所使用的方法包括两方面的 G-研究(见第 3 章),即特征和方法。多特征-多方法矩阵使我们把方差来源确定为“方法”和“特征”(或者“结构”)。因此,关于结构效度我们可以采用更精确的表述,因为这使我们把真正反映结构相似性的共变(并且因此与结构效度有关)与使用类似的测量程序而产生的人为的共变(因而与结构效度不相关)区别开来。当我们简单地考察两个测量之间的单一相关时,这种区别就不可能。

表面效度(face validity)

很多人使用表面效度这个术语来描述那些对他们表面上要测量的内容进行评估的一套题项。按照我的观点,由于以下几个原因,这种用法是不合适的:

首先,关于一个量表对其所测量的内容看起来像什么进行评价的假设可能是错误的。例如,艾德尔和本亚弥尼(Idler & Benyamini,1997)检验了 27 个大的、实施得很好的流行病学研究来精确地确定涉及了一个什么样的共同题项。该题项要求人们把他们的总体健康状况评价为“糟糕”、“一般”、“好”,或者是“非常好”。很多人都会判断出这个单一题项的测量确实评价了它所说的:被试的健康。艾德尔和本亚弥尼注意到,这个题项是对大量的健康结果的极好预测指标。在对不同的研究中的方差进行解释方面,它

总是超过了其他变量。与我们的讨论更有关的是，它初看起来似乎与健康状态不相关。然而，模型常常包括这个单一的题项并且也建立了关于健康状态的测量。典型地，单一题项健康自我评价和其他的健康状态测量在同一个模型中都是重要的预测指标。即是说，它们不共享来自其中一个的预测作用的足够方差，而排斥来自另外一个的独立的预测作用。相反，单一题项健康自我评价似乎与心理变量共享有更大程度的方差。这些发现表明，正如其表面上所表现的那样，这个被大量使用的单一题项不是健康状态的一个有效的指标。对于这个题项，看起来好象测量了我们想要的东西，但是对于支持效度来讲是不充分的。

以表面效度为基础来评价一个测量的另外一个问题是，有时候正在测量的变量的重要性并不明显。例如，一个打算测量人们的撒谎程度(例如，使他们自己“看起来很好”)的工具，由于很难使被试明白其意图，从而达不到其想要的结果。我们能够因为它看起来并不是在测量撒谎，而得出结论认为其是无效的吗？因此，这个例子表明，根据看起来无法知道它实际是什么，我们并不能得出它没有效度这一结论。

表面效度的最后一个问题是，量表的测量目的对谁来说应该是明显的，这还不清楚。是被试吗？如果一个医生问一个病人他或她是否比平常更口渴的话，那么这个问题的效度依赖于病人知道为什么要问这个问题吗？显然不是。是那个编制这个工具的人应该知道其目的吗？很难想象，对于量表的编制者来说，还不明白量表内容和所研究的变量之间的联系(或许，关于纯粹实验的、非理论的标准一相关效度除外)。如果这个意义上的表面效度被采用的话，实质上所有的量表都可以被判断为有效的。最后，从表面上来看，应该知道量表的测验目的人是一个更为广泛的科学团体吗？这种理解很可能产生有冲突的结论。在有些专家眼里，一个看起来像是测量了一个变量的题项，在另外一个相同资历的专家看来却测量了另外一个变量。而情况往往是这样的，那些根据一个量表似乎有或没有表面效度就认为其有或没有效度的人，只是根据自己个人的知觉来作出判断和认可。即是说，如果一个量表的目的和表面对他们来说看起来都相似，他们就倾向于认为其有

表面效度；否则，他们就认为其没有表面效度。这对于任何效度的宣布来说，其基础都是很脆弱的。

根据不同的环境，一个量表的目的从其表面明显地表现出来，可能有好处也有坏处。正如我们将在下一章中要看到的那样，题项编制过程经常需要明确地指向研究中的变量的陈述。这往往不是一件坏事。同时，我也并不是在说明测量工具总体上都应该这样来构建，以致它们的目的从其表面上看并不明显。相反，我想说明的是，无论是不是这样，结果都与效度几乎无关或根本无关。

练　习

1)举一个例子来说明，一个量表和一个行为之间的同等的相关如何表现结构效度或者是标准—相关效度。并对以下两个问题都进行解释：①计算相关的动机，以及②对这个相关的解释会随着研究者正试图评价的效度类型而变化。

2)假设研究者有关于两个结构：自信和社会适应的书面测量，研究者也有关于这两个结构的面试分数。在一个多特征—多方法矩阵中如何使用这些数据来证明：数据收集的方法对所取得的结果有一个不期望的强烈影响？

量表编制指南

Guidelines in Scale Development

步骤 1:清楚地决定你要测量什么

步骤 2:建立一个题项库

步骤 3:决定测量的模式

步骤 4:让专家评价最初的题项库

步骤 5:考虑确认题项的包含性

步骤 6:在一个试测样本中测试题项

步骤 7:求题项的值

步骤 8:优化量表长度

到目前为止，所呈现的材料都比较抽象。现在我们来看看怎么能应用这些知识。本章提供了一套具体的指导方针供研究者编制量表使用。

步骤1：清楚地决定你要测量什么

这一点看起来很简单。而且很多研究者认为，对于他们想要测量的内容早已有了一个清晰的想法，但是实际上却发现他们的想法比他们曾经认为的要含糊。通常，这种认识发生在为编写题项和收集数据进行了大量的工作之后——此时的改变，比在编写程序一开始时就意识到时成本要大得多。量表应该以理论为基础，还是应该探寻新的明智的方向？测量应该有多具体才好？某个现象的某些方面被强调时，其他方面应该被忽略吗？

有助于清晰化的理论

正如在第1章中所讨论的那样，清晰地思考量表的内容需要清晰地考虑所测量的结构。虽然在编制和验证一个量表中会涉及很多技术方面的问题，但是我们不应该忽略与测量的现象相关联的理论的重要性。在本书中关注的量表类型，是为了测量那些无法直接观察到的难以捉摸的现象。因为没有一个切实的标准让我们可以依此实施量表，有一些清晰的理论作为指导就很重要。同时，大量的现象必须进行再认，以防止量表的内容在无意中漂移到无关的领域中去了。

理论指导对于清晰化有很大的帮助。在编制本书中所讨论的量表时，也要考虑相关的社会科学理论。如果发现，现有的理论对量表编制者并没有任何指导的话，他们可能需要寻求一个新的明智的方向。然而，这个决定应该是一个以见多识广为基础，并且只有在对与目前的测量问题有关的合适的理论进行综述之后才做出决定。即使没有可以利用的理论来指导研究者，在编制量表之前，也必须制定出自己的概念性方案。从本质上讲，必须至少确定一种试验性的理论模型来指导量表的编制。这可能就如清楚地阐述

所要测量的现象一样简单。最好还应该包括一个关于新的量表结构与现有现象及其操作之间如何相关的描述。

有助于清晰化的特异性

所测量的结构的特异性或普遍性水平也很重要。在社会科学中有一个共识,即当变量与特异性水平相匹配的时候,它们之间将彼此高度相关(关于这点的讨论,参见 Ajzen & Fishbein, 1980)。有些时候一个量表的目的与非常特定的行为或结构有关,而另一些时候则需要寻求一个较普遍和一般的测量。

下面就举例来说明一下在特异性方面有区别的测量。考虑一下控制点(locus of control,LOC)结构。控制点是一个广泛使用的概念,指个体对影响他们取得重要成就的人物和事物(事件)的知觉。这个结构可以被广泛用作对在很多情境中的普遍行为的模式进行解释的一种方法,或者狭义地讲,用来预测个体在一个特定的情境中将会如何反应。关于影响的来源也被广泛地或具体地加以描述。例如,与这些理解一致,若特尔(Rotter,1966)的内部-外部(I-E)量表就是关注一个非常普遍的水平。其范围是从个人支配到由外界因素支配的一个单一维度,这也是这个量表的基础;并且其题项所强调的结果也是普遍性的,例如个体的成功。支配的外部来源也被广义地加以描述。以下就是若特尔的 I-E 量表中的一个外部陈述:"世界是由少数掌权的人所支配的,并且几乎没有哪一个小人物能够支配它。"勒温逊(Levenson,1973)编制了一个复合维度的 LOC 量表,它涉及三个控制点:自己,其他有权力的人,以及机会或命运。然而,她所关注的结果仍然是普遍的。以下是勒温逊的"掌权的其他人"这一分量表中的一个样本题项:"我感觉好像发生在我生活中的一切,都是由其他有权利的人所决定的。"沃斯顿和德维利斯(Wallston and DeVellis, 1978)使用勒温逊的三个控制点编制了多维健康控制点量表(multidimensional health locus of control, MHLC),其结果特指健康,例如避免疾病或生病。以下是 MHLC 中"其他有权利的人"这个分量表中的一个样本题项:"与我的医生经常保持接触,对我来说是避免疾病的最好方式。"沃斯顿、斯腾以及史密斯(Wallston,Stein,and Smith,1994)后来编制

了一个其结果更加特异性的多维健康控制点量表(MHLC 版本 C),它包括一系列“模板”题项。通过把每一个模板题项中的疾病或紊乱的名称替换成“我的状况”,该量表使研究者可以确定任何感兴趣的健康问题。以下是 MHLC 版本 C 中“其他有权利的人”这一分量表中的一个样本题项,它或许会用在糖尿病的研究中:“如果我经常去看我的医生,我的糖尿病出问题的机会就会更少。”

这些日益增多的更为具体化的 LOC 量表中的每一个都有潜在的用途。哪一个最有用,这取决于结果或控制的普遍性与所要探讨的特定问题之间的相关水平。例如,如果一个多维健康控制点量表的目的在于预测一个普遍的行为,或者会与其他在一个普遍的水平上来评价结构的变量相比较的话,那么若特尔的量表或许就是最好的选择,因为它也是普遍化的。另外一方面,如果研究者的兴趣在于具体地预测关于其他人的影响的信念如何影响某些健康行为的话,那么沃斯顿、斯腾以及史密斯(1994)量表可能更为恰当,因为特异性的水平与这个研究问题相匹配。在其编制过程中,就量表的目的性与功能而言,这些量表中的每一个都有一个清晰的框架来进行预测,该预测决定了哪种水平的特异性是合适的。关键在于量表编制者应当把这种思考当作一个积极的决定,而不仅仅是产生一套题项并且在犯了错误之后才看清楚它是什么样子。

关于控制点的例子证明了与结果(例如,这个世界是如何运作的以及糖尿病问题)和控制点(例如,普遍的外部因素与命运以及掌权的其他人)有关的特异性。然而,量表的特异性会随着大量的维度而变化,包括内容范畴(例如,焦虑与更广泛的心理调节)、背景(例如,特意编制来调查某个具体工作环境的问卷),以及人物(例如,儿童与成人或者军人与大学生)。

清晰地知道一个量表将包括哪些内容

量表编制者应该问问他们自己,他们想要测量的结构是否与其他结构有区别。正如先前所讲的那样,根据量表所应用的情境,量表可以被编制为相对广泛的或狭窄的。而对于它们所包括的结构,也是如此。测量普遍的焦虑是相当正统的。这种测量既应当

评估考试焦虑也应当评估社会焦虑。如果它与量表编制者或使用者的目标相匹配,那就更好。然而,如果我们只是对焦虑的某个特定类型感兴趣,那么这个量表应当排除其他的成分。那种"横漂"到一个相关的结构中去的题项(例如,当所感兴趣的主题是考试焦虑时,涉及了社会焦虑)是有问题的。

有些时候,明显地相似的题项会涉及完全不同的结构。在这种情况下,虽然量表的目的可能在于测量一个现象,但是它也会对其他现象很敏感。例如,某些抑郁量表,如流行病学研究中心(center for epidemiological studies depression, CES-D)抑郁量表(Radloff,1977),有一些涉及了抑郁的生理方面的题项(例如,与被试的"启动"能力有关)。在某些关于健康状况的情境中,例如关节炎,这些题项或许会把疾病的某些方面误认为是抑郁的症状(对于这一点的具体讨论,参见 Blalock, DeVellis, Brown, & Wallston, 1989)。如果量表被用于某些特定群体(例如慢性病)或者和其他一些生理结构一起使用(例如臆想病)的话,一个新的抑郁量表的编制者应当选择避免生理方面的题项。当然,如果是有其他用途的话,那么包括生理题项就很重要,例如当调查的范围尤其与具有负面影响的生理因素有关时。

步骤 2:建立一个题项库

一旦一个量表的目的被清晰地阐述明白了,编制者就会迫不及待地准备开始构建该工具。第一步就是要生成一个大的题项库,作为量表的最终候选题项。

选择反映量表目的的题项

显然,这些题项应当根据心目中的特定测量目标来选择或者编写。关于量表的实际意图的描述会引导这个过程。回想一下,我曾经提到过,组成一个相似的量表的所有题项应当反映作为其基础的潜在变量。就题项而言,每一个题项都被认为是一个对潜在变量的强度的"测试"。因此,每个题项的内容基本上都应当反

映问题中的结构。复合题项比单个题项所构成的测试更可靠，但是每一个题项仍然必须对潜在变量的实际分数敏感。

从理论上来讲，一套好的题项是从与问题中的结构有关的广泛的题项中随机选择而来的。题项集合被假设为无限的大，这就很大程度上排除了任何确切地识别它以及随机抽取题项的误差。然而，这种意识应当被放在心上。在编写新的题项，正如经常发生的情况一样，你应该创造性地思考一下你所寻求测量的结构。那么为了达到该结构，对一个题项进行措辞的方式是什么？虽然这些题项不应该在所定义的结构的界限以外去冒险，但是应当在这些范围之内寻找所有题项类型的可能性。一个量表的特征是由组成该量表的题项所决定的。如果它们是对你所一直持有的概念的一个较差的反映并且很难表达清楚的话，那么该量表就不会准确地把握所要测量的结构的实质。

题项所共同具有的"东西"确实是一个结构并且不仅仅是一个类别，这一点也很重要。再次回想一下，我们的量表模型把题项看作是作为其原因的共同潜在变量的明显证明。与一个共同的结构相关的题项的分数是由该结构的实际分数所决定的。然而，正如在第1章中所提到的那样，仅仅因为题项与一个共同的类别相关，这并不保证它们有相同的潜在变量。如态度、依从障碍，或者生活事件这些术语，经常定义的是结构的类别而非结构本身。举例来说，最终将会成为一个多维度量表的基础的题项库不应当仅仅关注态度，还应该关注特定的态度，例如关于惩罚毒品上瘾者的态度。如果你愿意，可以预想一下这个人的特征，即能造成对与惩罚该毒品上瘾者有关的题项做出反应的潜在变量。想象一个对普遍的态度进行解释的特征是一个相当大的挑战。对于所引用的其他例子情况也是一样。依从障碍在很多种类型中都比较典型。每一种类型（例如，发现症状的恐惧、对治疗成本的关注、对疼痛的预料、与治疗机构的距离、不会受伤害的感知）都可能代表一个潜在变量。在一些潜在变量中可能存在着非同寻常的相关。然而，这些障碍中的每一个都是一个单独的结构。因此，"障碍"这个术语描述了结构的一个类别而不是与单个的潜在变量相关的一个单独的结构。当它们是一个共同的潜在变量的表现时，测量同一类别

中的不同结构的题项（例如，不会受伤害的感知与对治疗成本的关注）就应该不会与题项以同样的方式共同变化。

冗 余

在量表编制过程的这一个阶段，如果其他一切都相同的话，最好是多包含一些题项。在编制一个量表时，冗余并不是一件坏事。实际上，那些指导我们的量表编制工作的理论模型就是以冗余为基础的。在讨论第 3 章中的斯皮尔曼—布朗预测公式时，我指出，如果所有其他条件都相等的话，那么信度作为题项的数量的一个函数而变化。通过编制一套以不同方式来反映某个现象的题项，我们才能试图把握问题中的这一现象。通过使用复合以及表面上冗余的题项，当它们的无关特质被删除掉以后，与这些题项都相同的内容就可以在整个题项中求和。如果没有冗余，以上操作就不可能。有用的冗余从属于结构，而不是题项的偶然性方面。在一个题项中仅仅把“一个”改为“这个”肯定都会带给你与题项的重要内容有关的冗余，但是它也可能是其他你所想要改变的东西的冗余，例如基本的语法结构以及词语选择。另外一方面，两个题项，例如“我会尽全力来保证我孩子的成功”与“如果它有助于我的孩子取得成功，再大的牺牲都不为过”，可能会是相当有用的冗余，因为它们以某种不同的方式表达了一个相似的思想。

与最后的量表相比，在你的题项库中，可以容忍有更多的冗余，并且某些冗余甚至在以后会更令人满意。例如，如果你有这样一个题项“就我的观点看来，宠物爱好者都很善良”，那么再包括另外一个表述“据我估计，宠物爱好者都很善良”的题项，就明显会更好一点。这些题项清楚地涉及了与宠物主人的身份有关的相似的句子，并且它们也有一个共同的语法结构并且使用几乎一样的词汇。然而，这样的一个题项“我认为喜欢宠物的那些人是善良的”，在与第一个题项的实质内容有关的冗余性方面会表现得很好——没有价值不高的冗余。然而在量表编制的这一较早阶段，即使是这个例子中的这两个最初的题项的极度冗余也是可以接受的，虽然只有一个会出现在最后的量表中。如果考虑两个题项，即使当它们像以上这些一样相似时，可能会为量表编制者提供一个机会

来比较它们，并且表达了一个偏好（例如，使用“观点”一词就似乎比“估计”一词要显得不那么自命不凡）。如果只考虑了这两个题项中的一个，那么这个机会就可能会失去。

题项的数量

对那些应该被包括进一个初步的题项库里的题项的数量进行详细地说明是不可能的。但是可以肯定地说，在最初的题项库中的题项，应该比你所计划的最终的量表中的题项要多得多。回忆一下，我们讲过，内部一致性信度是题项彼此之间（并且因此与潜在变量之间）的相关强度以及在一个量表中你所拥有的题项的数量的一个函数。由于题项之间的相关的本质在量表编制的该阶段还不清楚，有大量的题项是避免较差的内部一致性的一种保险形式。在你的题项库中的题项越多，在选择适合你的目的的题项方面就会越烦琐。在编制量表时，以一个其数量为最终量表的题项的 3 倍或 4 倍的题项库开始，这往往是非常平常的事情。因此，一个有 10 个题项的量表可能来自一个有 40 个题项的题项库。如果一个特定内容范畴的题项特别难以编写，或者如果实验数据表明无需大量的题项来取得好的内部一致性，那么最初的题项库比最终的量表大一倍就可以了。

总之，题项库越大越好。然而也有可能编制了一个太大的题项库，以致无法在所研究的问题的任何一个单一场合应用。如果题项库超乎寻常地大，研究者可以根据优先标准删除一些题项，例如缺乏清晰性、不恰当，或者与其他题项之间的不合意的相似性。

开始编写题项

开始编写题项往往是题项生成过程中最难的部分。让我来描述一下我是怎么开始这一过程的。在这一点上，我对题项的质量比对仅仅表达相关的意思的关注要少。我常常以一个陈述开始，即对我想要测量的结构的一个解释。例如，如果我对编制一个关于对商业信息的自我感受性的量表感兴趣的话，我会以这样的陈述开始“我容易受商业信息影响”。然后我会写出其他一些陈述，这些陈述有共同的意思但是在某些方面又有所不同。我的下一句

陈述可能会是,“商业信息对我有很多影响”。我会以这种方式继续,但是在这些陈述上没有施加任何实质上的质量标准。在这一早期阶段,我的目标仅仅是确定我期望的测量工具的中心概念得以表述的各种方式。当我在写的时候,我会寻求表达相似意思的其他方式。例如,我会在接下来的一系列句子中,用“我在电视或杂志广告上所看到的东西”来代替“商业信息”。我发现,快速而不加评论地写是很有用的。在写完了我预期包括在最终的量表中的题项的 3 倍或 4 倍题项时,我就检查一遍我所写的题项。现在就到了评价的时候了。可以根据把握中心意思的程度以及表达的清晰性来检查题项。接下来的阶段,在从原始的陈述列表中进行选择或对之进行修改的过程中,描述一下那些应该避免的特征或者应该合并的具体题项的特征。

好的题项和差的题项的特点

把造成一个题项所有好的或者坏的情况列出来是不可能的。内容范畴显然对题项的质量有重要的意义。然而,把握几个特点可以有效地把好的题项与差的题项分开。这些特点中的大多数都与清晰性有关。正如在第 1 章所指出的那样,一个好的题项应该是不含糊的。那些使被试进退两难的问题应该被删除。

量表编制者应该避免编写太长的题项,因为长度往往会增加复杂性面降低清晰性。然而为了简洁而牺牲题项的意思是不可取的。如果一个正在修改的句子对于转达一个题项的意图是关键的,那么就应把它包括进来。然而,要避免不必要的华丽辞藻。总之,像这样的题项“我经常在阐述论点方面有问题”会比一个不必要的较长的题项要好,例如“公平地讲,很多时候我似乎都有一个问题,就是让其他人理解我的论点。”

在选择和编制题项时,另外一个需要考虑的问题是,所编写的题项的阅读难度水平。有大量的方法(例如,Dale & Chall, 1948; Fry, 1977)来评估文章段落的等级水平,包括量表的题项。这些方法都把较长的词汇和句子等同于较高的阅读水平。阅读大部分地方报纸可能需要一个六级阅读水平。

弗瑞(Fry,1977)描绘了量化阅读水平的几个步骤。首先是选

择一个课本范例,该范例以一个句子的第一个单词开头并刚好只有 100 个单词(对于那些只有少数几个题项的量表,你或许必须选择 100 这个方便的数字,并以此作为后续步骤的基础)。其次,计算在这个课本范例中的完整句子和单个音节的数量。这些数值被用作图表的记分点,该图表为这 100 个单词的段落的不同句子组合和音节数提供了等级等价物。这个图表明,对于一个阅读水平为第五级的句子来说,每个句子中的单词的平均数量和音节的平均数分别是 14 和 18。在第六个等级水平的一个句子中,平均有 15 或 16 个单词以及总共 20 个音节;在第七个等级水平的一个句子中,有大约 18 个单词和 24 个音节。有较多的较长的单词的较短的句子,与具有较少的长单词的长句子相比,会得出一个相同的等级水平。例如,一个有 9 个单词和 13 个音节的句子(例如,有 44%的多音节单词)以及一个有 19 个单词和 22 个音节的句子(例如,不到大约 14 的多音节单词)都被归类为第六个等级的阅读水平。对很多用于普遍人群的测量工具来说,把阅读水平确定在第五和第七个等级之间可能是一个合适的目标。例如,多维健康控制点量表的题项,是在第五到第七等级的阅读水平上编写的。在这个阅读水平上的一个典型的题项是:"影响我健康的大多数东西都是偶然发生的"(Most things that affect my health happen to me by accident; Wallston et al.,1978)。其有 9 个单词和 15 个音节,把它放在了第六个等级水平上。

弗瑞(1977)阐述道,在评价阅读难度时,应该考虑语义和句法因素。因为短的单词似乎更普遍,并且短的句子在句法上似乎也要简单些,所以该程序是其他更复杂的难度评价方法的一个可接受的替代方法。然而,当使用其他标准来编写题项和选择题项时,在应用阅读水平方法时我们必须使用一定的常识。仅仅包含短单词的一些简洁的句子并不是基本的。例如,对有些受过学校教育的人来说,"避开头盔的嘲笑"(Eschew casque scorn),可能会比"戴上你的头盔"(Wear your helmet)更令他们感到迷惑,尽管它们都有 3 个单词和 4 个音节。另外一个应该避免的潜在的混淆因素是多重否定:"我不赞成对那些反核武器的组织停止资助的合作"就比"我赞成继续对那些宣传禁止核武器的群体提供私人的支持"要

令人困惑得多(注意观察,这两个陈述可能表达了在同一问题上的不同立场。例如,后者可能隐含着对问题中的群体提供私人支持而非公共支持)。

所谓的"双筒枪"(double barreled)题项也应该被避免。这些题项传达了两个或更多的意思,因此对这种题项的认可,有可能指两者之一,也有可能指二者。"我支持公民权利,因为歧视是对上帝的一种犯罪"就是一个双筒枪题项的例子。如果一个人支持公民权利的原因在于其他,而不是在于对神的蔑视的话(例如,因为它是对人性的一种罪行),那么他或她将怎么回答?一个否定的回答或许会不正确地传达一个缺乏对公民权利的支持,而一个肯定的回答可能不正确地把动机归结为被试的支持。

量表编制者应该避免的另外一个问题是模棱两可的代词。"谋杀犯和强奸犯不应该从政治家那里寻求庇护,因为他们是地球的浮渣"(Murderers and rapists should not seek pardons from politicians because they are the scum of the earth),只要不考虑代词的指代,这句话可能表达了某些人的观点(然而,量表编制者通常更清楚一个题项的意义是什么)。这个句子应当受到双重批判。除了其模糊的代词以外,它也是双筒的。位置不当的修饰语也产生与模棱两可的指代相似的模糊性:"我们众议院中的议员应该努力工作来为卖淫方法"(Our representatives should work diligenty legalize prostitution in the House of Representation)就是这种修饰语的一个例子。考虑一下这两个之间在意思上的差别:"所有的游民都应该被给予精神分裂症的测评"(All vagrants should be given a schizophrenic assessment)测评"与"所有的游民都应该被给予精神分裂症测评"(All vagrants should be given a schizophrenia assessment)。

单个的单词并不是暧昧性的惟一来源。整个句子可能有不止一个意思。我曾经看到过一个关于青少年性行为的调查,该调查包含一个父母的教育评价。把调查的上下文作为一个整体来看,其措辞却非常不幸:"你的母亲在学校里走了多远?"(How far did your mother go in school)研究者始终没有认识到这个陈述的无关意义,直到在一次学术会议上做报告时它被一个专家组所窃笑时

才发现。我怀疑这个题项也被大量的被试讥笑了。还不清楚它将怎么样影响青少年对问卷的后续部分的反应。

措辞积极和消极的题项

很多量表编制者选择编写措辞消极的题项，即表示低水平的或者兴趣结构缺失的题项，以及更普遍的代表该结构存在的措辞积极的题项。之所以这样的目的在于得出一套题项，既有被认可时反应出的高水平潜在变量的题项，也有不被认可时反应出的高水平变量的题项。例如，罗森伯格自我评价量表(RSE,Rosenberg,1965)包括表示高自我评价的题项(例如，“我感觉我有很多好的品质”)和低自我评价题项(例如，“我确实不时感到很无用”)。在同一个量表中对题项的措辞既有积极的也有消极的，其目的往往是为了避免默许、断言，或者一致性偏见。这些内部变化的题项使被试的反应倾向与题项一致，而不管其内容。例如，如果一个量表由表示高水平的自我评价的题项所组成的话，那么默许性偏见会导致其反应模式似乎表示非常高的自我评价。另外一方面，如果量表是由数量相等的积极措辞和消极措辞的题项所组成的，那么默许性偏见和极度的自我评价可以通过反应模式而相互区别。一个“中性者”既可以认可表示高自我评价的题项也可以认可表示低自我评价的题项，而一个真正有高自我评价的人会始终认可高自我评价题项并且消极地认可低自我评价题项。

不幸的是，既包括积极的题项也包括消极的题项可能会付出一定代价。题项极性的逆转可能会对被试造成混淆，尤其是当他们在完成一个长的问卷时。在这种情况下，不管其极性，被试可能会对表达他们对一个陈述的强烈同意，与表达他们对所测量的属性(例如，自我评价)的强度之间的差别感到困惑。作为一个应用社会科学的研究者，我曾经看过很多以相反的方向措辞但是编写得很差的题项的例子。例如，在德维利斯和卡拉罕(DeVellis & Callahan, 1993)的著作中，我的同事和我描述了一个更简短的、更聚焦的术语来替代“风湿病学态度索引”(rheumatology attitude index，一个很不幸的名称，因为该工具并没有评价态度并且也不是一个索引)。我们从以实验标准为基础的原始的较长版本中选择

题项，并且以四个表示对疾病的消极反应的题项、一个表示很好地应对这种疾病的能力的题项结尾。目的在于，使用者应该对这个"应对"题项反向打分，以便所有的题项都表达了一种无助感。更近一些的研究，克尔瑞、卡拉罕和德维利斯（Currey, Callahan & DeVellis, 2002）检验了那个惟一的以积极方向措辞的题项，发现该题项的正确率总是很差。当该题项被增加了一个否定词"不"来改变其分值，以便和其他的题项一致时，其正确率就显著地提高了。

我们怀疑，虽然很多被试认识到了最初的题项的不同分值，但是其他人并不一定意识到。这就会导致一部分人会认为最初的题项与其他题项有积极的相关，而另外一部分却认为这种相关是消极的。结果，对于作为一个整体的该例子来说，该题项与其他四个题项之间的相关将会显著地被减少，并且因此产生我们所观察到的在最初的题项（反向评价）上的不令人满意的分数。以社团为基础的测试样本的个人经验告诉我，以相反方向措辞的题项有百害而无一利。

小 结

一个题项库应该是一个量表得以形成的丰富资源。它应当包含大量的与研究兴趣有关的题项。与内容有关的冗余是一笔资产，而不是一笔债务。它是内部一致性信度的基础，反过来信度又是效度的基础。题项不应该涉及"一揽子交易"，因为它会使被试不可能认可题项的一部分而不认可与该部分不一致的其他部分。不管在题项库中是否都包括了积极地和消极地措辞的题项，它们的措辞都应当遵循已建立的语法规则。这将有助于避免以上所讨论的导致模糊性的一些因素。

步骤 3：决定测量的模式

有着大量的测量模式可供选择。研究者应该提早考虑测量的模式是什么。这个步骤应该与题项的编写同时进行，以便二者的

一致。例如,如果最终所选择的反应形式是由单个词语题项所组成的一个题项集,那么生成一个长的由陈述句组成的题项集就可能是浪费时间。此外,相比而言,预设的理论模型会与某些反应形式更一致。总之,由可以连续记分以及求和,从而获得一个分数的量表,与本章所提出的理论倾向非常一致。然而,在本节中,我将讨论一些普遍的测量模式,这些模式以第 2 章中所讨论的理论模型所隐含的形式为基础。

瑟斯顿测量(thurstone scaling)

对于编制影响题项形式和反应选择的量表,存在有大量的常规性策略,一种方法就是瑟斯顿测量,这里有一个类比可以帮助阐明瑟斯顿测量是如何工作的。设计一把转动的叉子并且以一种特定的频率振动。如果你敲击它,它就会以那种频率振动,并且产生一种特定的音调。相反,如果你把这把叉子放在接近与这把转动的叉子所产生的频率一样的音调源的地方,这把叉子也将开始振动。那么,从某种意义上说,一把转动的叉子就是一个"频率检测器"。它会因为其共振频率的声波的出现而振动,而当出现其他频率时却保持不动。假想一下,一系列振动的叉子排成一行,以致当一把叉子从左到右移动时,这些振动的叉子就会相应地产生逐渐增高的频率声音。在这些振动的叉子的频率范围之内,这种排列方式就可以用来确定一个音调的频率。换句话说,你可以通过观察当该音调出现时,哪一把叉子会振动,从而确定该音调的频率。瑟斯顿测量就是以这样的工作方式来编制的。量表的编制者试图编制一些对问题中的特征的不同水平作出不同反应的题项。当某个特定题项的"音调"与被试所具有的特征水平相匹配时,题项将发出"信号"来表示这种一致性。通常,"信号"由一个对与该特征的合适水平"和调"的系列题项的肯定反应和一个对所有其他题项的否定反应所组成。典型地,通过把大量的题项放入与结构强度或力度的等距对应的题项集中,从而做出判断来检测该调谐(例如,决定每个题项所对应的结构的水平)。

这真是一个一流的想法。例如,可以编制出与某个特征的不同强度相对应的题项,可以留一些距离来表示相等的间隔,并且可

以格式化来对应赞同—反对反应选择。研究者可以把这些题项给被试，然后检查他们的反应来确定哪个题项引发了一致性。因为可以预先校准这些题项对某个现象的特定水平的敏感性，所以这种一致性可以确定该被试所拥有的特征的数量。选择题项来表示题项间的相等间隔会得到高度满意的测量特征，因为分数可以服从以间距测量为基础的数学程序。

以下是用来测量父母对其孩子的教育和职业成就期望的瑟斯顿量表的一部分：

(1)对于我的孩子来说，取得成功是对作为一个父母的我所付出的努力回报的惟一方式。	同　意______ 不同意______
(2)上一所好的大学并且找到一个好的工作非常重要，但是对我的孩子的幸福却不是至关重要的。	同　意______ 不同意______
(3)幸福与取得教育或物质目标无关。	同　意______ 不同意______
(4)传统价值观中有关成功的诱惑力是对真正幸福的一个障碍。	同　意______ 不同意______

正如农纳利(Nunnally,1978)所指出的那样，编制一个真正的瑟斯顿量表比描述一个瑟斯顿量表要困难得多。找到与所研究的现象的特定水平相一致地“共振”的题项十分困难。该方法所遇到的实际问题经常超过了其优越性——除非研究者有一个迫不得已的原因，从而需要它所提供的校准类型。虽然瑟斯顿量表是一个令人感兴趣的并且有时也是合适的方法，但是在以后的章节中我将不会再对它做讨论。然而，请注意，当采用某种不同的方法来测量它们的时候，在第 7 章中所讨论的以项目反应理论为基础的那些方法，拥有瑟斯顿量表的很多成分。

加特曼测量(guttman scaling)

加特曼量表是由一系列题项所组成的，这些题项涉及一个特征的逐渐升高的不同水平。因此，被试应该认可大量相近的题项，直到在某一个关键点，这些题项所涉及的特征的总量超过了被试所拥有的特征总量为止，剩余的所有题项都不会被认可。有些纯描述性数据适合加特曼量表。例如，很多面试问题可能会这样问，“你抽烟吗?”“你每天抽烟超过 10 支吗?”“你每天抽烟超过一盒吗?”等等。就这个例子而言，认可加特曼量表中的任何特定题项就意味着对其前面所有题项的认同。被试在该特征上的水平可以通过对最高一个题项的赞同反应来确定。请注意，尽管瑟斯顿量表和加特曼量表都是由等级化的题项所组成的，但是前者的焦点在于一个单独的肯定反应，而从肯定反应过渡到否定反应这一点则是后者的焦点。对于先前所例举的父母的期望量表，其加特曼版本可能像这样：

(1)对于我的孩子来说，取得成功是对作为一个父母的我所付出的努力回报的惟一方式。	同　意______ 不同意______
(2)上一所好的大学并且找到一个好的工作对我的孩子的幸福是非常重要的。	同　意______ 不同意______
(3)如果一个人已经取得了他或她的教育或物质目标，幸福就更可能实现。	同　意______ 不同意______
(4)传统价值观中，关于成功的诱惑力并不是对真正的幸福的一个障碍。	同　意______ 不同意______

加特曼量表对于客观的信息或情境处理得非常好，因为积极地对一个等级的一个水平作出反应就意味着满足了该等级的所有较低水平的标准，这在逻辑上是必要的。当所研究的问题并不具

体时，情况就变得比较悲观。例如，在我们所假设的父母的期望量表这一例子中，在每个个体之间的等级可能不统一。尽管每天抽 20 支香烟总是表明要比抽 10 支多，但是在父母的期望量表这一例子中，对题项 3 和 4 的反应可能并不总是与加特曼量表的等级模式相一致。例如，一个人可能会同意题项 3，但是不同意题项 4。一般而言，赞同题项 3 就意味着赞同题项 4，但是如果一个被试把成功看作是一个复杂的因素，该因素同时作为对幸福的一个帮助或一个障碍，那么就会得到一个非典型的反应模式。

与瑟斯顿量表一样，加特曼量表无疑有其自身的价值与地位，但是它们的应用性看起来非常有限。对于这两种方法来说，其劣势和难度都将超过它们的优势。记住，目前所讨论的测量理论并不总是可以应用于以上类型的量表，这是非常重要的。当然，对潜在变量和每个题项之间的相同强度的因果关系的假设或许不能应用于瑟斯顿量表或加特曼量表。农纳利和博恩斯腾(Nunnally & Bernstein，1994)简要地描述了作为这些量表的基础的一些概念上的模型。对于尤其适合等级化题项的测量情境，在第 7 章中将要讨论的以 IRT 为基础的模型，是一个潜在的合适选择，尽管实施这些方法非常烦琐。

具有相同权重的题项的量表

先前所讨论的测量模型最适合由这样的题项所组成的量表，即对研究的现象来说是更加或较不相等的"指标"的题项——也就是说，它们是更加平行的或更不平行的(但是并不必要像平行测试模型那样严格意义上的平行)。它们是某个共同现象的非理想的指标，该现象可以通过简单的合计而被复合为一个可以接受的可靠的量表。

这种类型量表一个引人注目的特征是，每个单独的题项都能够有大量的反应选择模式。这使得量表编制者在编制一个最适合特定目的的量表时，有大量的纬度可以选择。下面将要讨论与反应形式有关的某些一般的问题，同时也将要讨论一些有代表性的反应模式的优点和要求。

反应类别的最适宜数量是多少

大多数量表题项包括两个部分:题干和一系列反应选项。例如,每个题项的题干可能是表达一个观点的一个不同的陈述句,而与每个题干相伴随的反应选项可能是一系列表示对该陈述句的赞同程度的描述。现在,让我们集中讨论一下反应选项——尤其是,那些可供被试所使用的选择的数量。一些题项反应模式提供给被试一个无限的或大量的选择,而其他题项反应模式则限制可能的反应。例如,假设有一个类似温度计一样的量表用来测量愤怒,从温度计的底部"根本就没有愤怒"校准到其顶部"完全的,无法控制的愤怒"。被试应该被呈示一系列的情境描述,每个描述都伴随一个温度计量表复印图片,并要求他们通过把温度计的某些部位涂上阴影来表示该情境激发了多大程度的愤怒。实质上这种方法允许对愤怒的连续测量。另外一种方法可能会要求被试用从 1 到 100 的数字来表示每个情境所激发的愤怒程度。这就提供了大量的离散的反应。此外,该形式可以把反应选择限制在少数几个选项上,例如"没有"、"有一点"、"中等量"和"很多",或者限制在"愤怒"与"不愤怒"之间的一个简单的二元选择。

这些不同方法的相对优势是什么?一个量表的令人满意的质量是可变性。如果它不变化的话,一个量表就不能共变。如果一个量表在潜在的特征之中无法区别方差的话,那么它与其他量表的相关将会受到限制并且它的应用也会受到限制。增加可变性的机会的一种方法是,拥有大量的量表题项。另外一种方法是题项内有大量的反应选项。例如,如果研究环境把研究者限制在与愤怒有关的两个问题上,最好在描述被试的愤怒水平中给他们以纬度。假设研究涉及在工作场合禁止吸烟的政策。让我们进一步假设研究者想考察该政策与愤怒之间的关系。如果研究者有两个开放式问题(例如,"当你被限制吸烟的时候,你会感到有多愤怒?"以及"当你遇到别人在工作场合吸烟的时候,你会感到有多愤怒?"),与二元选择形式相比,被试可能会从这个反应形式中获得更多有用的信息,因为这给了他们很多反应等级。例如,一个从 0 到 100 的量表在对这些情境的反应中可能会出现广泛的差异,并且为双

题项的量表产生了好的可变性。另一方面，如果条件允许，研究组有关于吸烟和愤怒的50个问题，通过增加题项来取得一个量表分数时，简单的"愤怒"与"不愤怒"指示可能会产生足够的可变性。事实上，如果被试在这50个问题的每个上都面临着更多的选择，那么这可能使他们感到疲劳或厌烦，从而降低他们的反应信度。

与反应选项的数量相关的另外一个话题是，被试有意义地识别的能力。有代表性的被试能够对这些选项作出多好的区分？这明显取决于所要测量的问题。几乎很少有东西能够被评价为50个离散的类别。当被呈现如此多选项时，很多被试只使用那些与5或10的倍数相对应的选项，从而有效地把选择的数量减少到5个。35和37这样的差别，可能无法真正反应所测量现象中的差别。虽然量表的方差可能会增长，但是它可能是由正在增长的潜在现象所引起的随机(例如，误差)部分而不是系统部分。这当然不会带来任何好处。

有些时候，被试有意义地在反应选项间进行区别的能力将取决于那些选项的特定措辞或者位置。要求一个被试对一些模糊数量的描述进行区别，例如"几个"、"少数几个"以及"很多"，会产生很多问题。有时候这种模糊性可以通过反应选项在页面上的安排而减少。当他们被呈现一个明显的连续统一体时，被试似乎总是能理解主试的意图。因此这样的一个序列，如：

很多　　一些　　少数几个　　极少数　　没有

可能暗示着"一些"要比"少数几个"要多。然而，如果有可能找到一个没有模糊性的形容词来排除，被试根据一个连续统一体的位置而做的假设的话，那就更好。有时候，较少的反应选项比模糊性的选项可能要好。因此，例如，在以上例子中把"一些"或者"少数几个"删除从而只有四个选项而不是五个，这可能会更好。最糟糕的情况是把模糊的词语和模糊的位置混杂在一起。看看以下这个例子：

非常有帮助	不是非常有帮助
有某些帮助	根本没有帮助

像"某些"和"不是非常"这样的词语在最好的情境下也是很难区分的。然而,如上面所列举的例子那样,如此排列这些选项会使情况变得更糟糕。如果被试先从第一列往下看,然后再从第二列往下看,"某些"似乎表示比"不是非常"的值要高。但是,如果被试从左到右从第一行看到第二行的话,沿着该连续统一体的两个描述的隐含等级就与前面刚刚相反。由于既有语言上的模糊性又有空间排列上的模糊性,个体会赋予这两个表示中等值的选项以不同的意义,因此其信度就会受到影响。

此外,还有一个问题需要考虑,那就是研究者对每个题项进行记分的能力与意愿。如果先前所描述的温度计方法是被用来量化愤怒反应的话,那么研究者真正打算企图对每个反应进行一个精确的记分吗?共同的领域能够被测量到四分之一英寸那么精确吗?四分之一厘米呢?四分之一毫米呢?如果从量表中提取出来的只有一些粗糙的数据,即低、中、高三等,那么在要求这样一个精确的反应中所需要注意的是什么?

至少还有一个问题与反应选项的数量有关。假设每个题项有一些离散反应,那么这个反应数量应该是偶数还是奇数?再次,这个问题取决于问题的类型、反应选项的类型,以及研究者的目的。如果反应选项是双极的,用一个极端来表示另外一个的反面(例如,一个强的积极态度与一个强的消极态度),那么奇数的反应选项允许模棱两可(例如,"既赞成也不赞成")或不确定(例如,"不肯定");偶数通常不会。奇数的反应选项意味着存在有一个中心的"中立"点(例如,既不是积极的赞同也不是消极的赞同)。相反,偶数的反应选项迫使被试在一个极端或另外一个极端的方向中至少要做出一个弱的许可(作为最不极端的反应,例如,在中等的积极赞同或中等的消极赞同之间做一个迫选)。任何一种形式都有其存在的必要性和相对的优势。如果感觉到被试将选择一个中立的反应来作为避免一个选择的方式的话,那么研究者或许会排除模棱两可的选项。例如,在关于社会比较性选择的研究中,研究者或

许会想迫使被试对关于一个更优越的人或较不优越的人的信息表示一种偏好。考虑一下这两种不同的形式，其中第一种是被用来研究患有风湿病的病人之中的社会性比较的（DeVellis et al.，1990）：

(1)你愿意听到关于以下哪个人的信息：

a. 那些患有风湿病且比你的病情更严重的病人

b. 那些患有风湿病且比你的病情较缓和的病人

(2)你愿意听到关于以下哪个人的信息：

a. 那些患有风湿病且比你的病情更严重的病人

b. 那些患有风湿病且和你的病情一样糟糕的病人

c. 那些患有风湿病且比你的病情较缓和的病人

像(2)b 那样的一个中立选项可能会引发不必要的含混。有时候，一个中立点或许也是必要的。在关于评价两项冒险行为中（例如，厌烦或愤怒）人们更喜欢哪个的一项研究中，一个中点可能很关键。在关于一个安全而沉闷的活动与一个兴奋而危险的活动之间的选择中，研究者或许应该变化损害的几率和程度。在关于对更兴奋的活动进行冒险的选项中，一个被试所选择的非常接近含混的那一点，就可以被看作是冒险行为的一个指标。

以下是关于活动 A 和活动 B 描述，请在活动 B 的选项中圈画出你认为适合你的选项，从而表示你对以下所列举的选项中对活动 A 或活动 B 的偏好：

活动 A：阅读一本统计书（没有受严重伤害的危机）

(1)活动 B：乘坐一架计算机化的小型飞机飞行（非常小的受严重伤害的机会）

强烈地　中等程度地　没有　强烈地　中等程度地

喜欢 A　喜欢 A　偏好　喜欢 B　喜欢 B

(2)活动 B：乘坐一架开放式座舱的小型飞机飞行（小的受严重伤害的机会）

强烈地　中等程度地　没有　强烈地　中等程度地

喜欢 A　喜欢 A　　偏好　喜欢 B　喜欢 B

(3)活动 B:从一架配有支撑性降落伞的飞机上跳伞(中等受严重伤害的机会)

强烈地　中等程度地　没有　强烈地　中等程度地

喜欢 A　喜欢 A　　偏好　喜欢 B　喜欢 B

(4)活动 B:从一架配无支撑性降落伞的飞机上跳伞(受严重伤害的机会很多)

强烈地　中等程度地　没有　强烈地　中等程度地

喜欢 A　喜欢 A　　偏好　喜欢 B　喜欢 B

(5)活动 B:从一架配有支撑性降落伞的飞机上跳伞,并且尝试在一个目标上着陆(总是肯定会严重受伤)

强烈地　中等程度地　没有　强烈地　中等程度地

喜欢 A　喜欢 A　　偏好　喜欢 B　喜欢 B

除开这种方法的优点或信度,它会明确地要求反应选项包括一个中点。

反应形式的具体类型

量表题项有各种令人头晕眼花的形式。然而,有几个方法可以表示那些被广泛使用的题项,并且在大量的应用实践中被证明是成功的。以下将讨论其中一些方法。

利克尔特量表(likert scale)

最一般的题项形式之一是利克尔特量表。当使用利克尔特量表时,题项以一个陈述句的方式呈现,后面跟着反应选项,表示对该陈述的赞同或认可程度的变化(实际上,前面那个选择冒险活动的例子就使用了利克尔特量表的反应形式)。根据正在研究的现象和研究者的目的,跟随每个陈述的反应选项的数量要么是奇数要么是偶数。对于该陈述来说,反应选项的措辞应该使其都有大致相等的间距。即是说,任何两个相邻的反应选项之间在赞同方面的差别应该大致与任何其他相邻的反应选项对之间的差别相同。一般的惯例是包括六个可能的选项:“强烈地反对”、“中等程

度地反对”、“稍微反对”、“稍微赞同”、“中等程度地赞同”、“强烈地赞同”。这些选项就形成了一个从强烈地反对到强烈地赞同的连续体。也可以加入一个中立的中点。对于这个中点，一般的选择包括“既不反对也不赞同”以及“反对与赞同的程度一样”。我们抽一点时间来讨论一下这两个中点的同等性。第一个暗示着无动于衷的缺乏兴趣，而后者表明对于赞同与反对都有一个强烈而相等的兴趣。最好是使大多数被试不是注意语言的微妙，而是仅仅把在中心范围的任何合理的反应选项都看作是一个与其精确的措辞无关的中点。

利克尔特量表在测量观点、信念和态度的工具中被广泛使用。对于这些陈述来说，当被用在利克尔特量表中时，恰当的（虽然不是特别地）强度经常是非常有用的。假设，在反应选项的选择中表示了适度的观念。例如，这些陈述“医生普遍地忽略病人的说话内容”、“有时，医生并没有给予他们本应该给予病人那么多的注意”以及“偶尔，医生可能会忘记或者疏忽病人所告诉他们的情况”分别表达了关于医生对病人的谈话的忽略的强烈的、中等的和弱的观点。对于利克尔特量表来说，哪个是最好的？当然，根本上最精确地反应了观点之间的实际差别的那个就是最好的。在最初的题项库中，在选择以何等强度对题项措辞时，研究者可以从对这样几个问题的回答中受益，“对于所研究的问题，有不同数量或强度的态度的人们可能会怎么反应?”在以上所举的三个例子中，研究者可能会得出结论认为最后一个问题可能表明了人们较强的赞同，这些人的观点遵循从肯定到否定的一个连续体。如果该观点被证明是正确的，那么第三个陈述就无法很好地区别有强烈的否定观点与中等程度的否定观点的人们。

总之，当用利克尔特量表中时，过于适度的陈述可能会引出太多的赞同。很多人会强烈地赞同像这样的陈述“市民的安全和保障非常重要”，人们可能会强烈地赞同这样一个陈述（例如，选择一个极端的反应选项）但却没有持有一个极端的观点。当然，反过来也是如此。那些没有持有最极端的观点的人可能会发现他们不赞同一个极端强烈的陈述（例如，“极力抓捕和惩罚罪犯比保护个人的权利更重要”）。在这两个（极度中庸或极度极端）陈述中，基于

两个原因，前者可能存有更大的问题。第一，我们经常偏爱书写一些不会冒犯我们的被试的陈述。避免冒犯或许是一个好主意，然而，它会使我们偏好那些几乎每个人都可能赞同的题项。对太中庸的题项保持警惕的另外一个原因是，它们可能会表示信念或观点的缺乏。前一段中所举的关于那个不专心的医生的题项中的第三个题项，并不表示一个赞同的态度，也不表示一个不赞同的态度。像这样的题项可能与研究目的很不相符，因为我们经常感兴趣的是某种现象的出现而不是其缺乏。

总之，一个好的利克尔特题项应该以清楚的术语来陈述研究中的观点、态度、信念或其他结构。对于这种类型的量表，超越关于这些结构的从弱到强的选项范围，既没有必要也不合适。反应选项提供了划分等级的机会。

以下就是在利克尔特反应形式中的题项的两个例子：

(1)锻炼是健康生活方式的一个基本成分。

1	2	3	4	5	6
强烈地反对	中等程度地反对	稍微反对	稍微赞同	中等程度地赞同	强烈地赞同

(2)禁毒战役应该享有国家级的优先权。

1	2	3	4	5
完全正确	大部分正确	同等地正确与不正确	大部分不正确	完全不正确

语义微分(semantic differential)

语义微分测量方法主要是与奥斯古德和他的同事(Osgood & Tannenbaum，1995)的关于态度的研究相联系的。一般，语义微分在指一个或多个刺激中使用。例如，在关于态度的情况中，刺激可能就是像汽车销售人员这样的一组人。在一系列配对的形容词之后跟随着对目标刺激的确定。每个配对代表着一个连续体的相反的两端，用形容词来定义(例如，诚实与不诚实)。就如以下例子所示，在组成反应选项的形容词之间有很多条线：

汽车销售人员

诚实 ____ ____ ____ ____ ____ ____ ____ 不诚实

安静 ____ ____ ____ ____ ____ ____ ____ 嘈杂

本质上，每一条线（7 和 9 是通常所使用的线条的数量）代表由形容词所定义的连续体中的一个点。被试在其中一支线条上做记号以表示连续体上的一个点，从而表示刺激的特征。例如，如果某个人认为汽车销售人员极其不诚实，他或她可能会选择最靠近那个形容词的那条线。无论是极端的还是中庸的观点都能够通过选择一条线并做上记号来表示。在完成与第一个形容词配对有关的刺激的等级划分之后，这个人就接着完成其他由线条所分离的形容词配对。

一个人所选择的形容词既可以是双极的也可以是单极的，这总是取决于量表所要测量的研究问题的逻辑性。双极形容词每一个都表达了相反的态度的出现，例如友好与敌意。单极形容词配对表示一个单独的态度的出现和缺乏，例如友好与不友好。

像利克尔特量表一样，语义微分反应形式能够与在本书中的前一些章节中所出现的理论模型高度一致。能够编写涉及相同的潜在变量的题项集合。例如，使用像“值得信赖/不值得信赖”、“公平/不公平”以及“诚实/不诚实”这样的词语来作为端点的题项，可以加到前面一个例子的第一个陈述中，从而组成一个测量“诚实”的量表。这样的量表能够被概念化为一套题项，这些题项享有一个共同的潜在变量（诚实）并且与第 2 章中所讨论的假设相一致。与之相对应，对于题项的评价，每个“诚实”题项的分数应该被求和，并且像在稍后一节中所要讨论的那样进行分析。

形象化类比（visual analog）

另外一个在某些方面与语义微分相类似的题项形式是形象化类比量表。这种反应形式向被试呈现一根连续的线条，该线条连接表示一个连续体的相反两端的一对描述。指导完成该题项的被试在线上做一个标记来表示他们的观点、体验、信念或者其他任何被测量的东西。形象化类比量表，正如其名称中的“类比”这个词语所表明的那样，是一个连续量表。在评价量表里所标记的点的

分数方面的差异好坏，是由研究者决定的。连续的反应形式的一些优点和缺点在先前已经讨论过了。那个时候还没有出现的一个额外的问题是，当其与这个连续体的评价有关时，在对物理空间的理解中可能存在差异。对不同的人来讲，沿着这条线在特定的点所做的记号可能并不意味着相同的意义，甚至当所有的被试在这条线上所标注的终点都是相同的时候，也是如此。考虑一下这样一个关于疼痛的形象化类比量表：

根本就没有疼痛	______________________________	我经历过的最严重的疼痛

在该量表中间的一个反应表示的是一半的时间上的疼痛，还是一半的程度上的连续疼痛，或者完全是其他的东西？关于疼痛的测量，部分问题在于疼痛可以在很多维度上被评价，包括频率、程度以及持久性。测量一个人所经历的最严重的疼痛可能会被曲解。不同个体间的对比也由于以上问题而变得复杂，不同的人可能经历的“最严重的疼痛”的水平不一样。当然，这些问题中的有一些存在于这个例子所使用的研究现象之中——疼痛（关于疼痛的测量的完整讨论，参见 Keefe，2000），而本质上并不存在于量表之中。然而，形象化类比量表中分值的异质性分配问题也会由于其他现象的影响而存在。

形象化类比量表的一个主要优点在于，它们非常敏感（Mayer，1978）。这能够使它们在一些干预事件之前和之后对研究现象进行测量非常有用，例如产生一个相对弱的影响的干预或实验处理。例如，在实验处理过程中的一个温和的责备对自信的 5 点测量不会产生一个转换。然而，向在形象化类比量表中的分值较低的一个突然而系统的转换，可能会发生在这个假设实验中的“责备”情境中的人身上。当对同一个体间的而不是不同个体间的随着时间的变化而发生的改变进行测量时，这种敏感性可能更优越（Mayer，1978）。事实可能的确如此，因为在前一种情况中，并不存在由于个体间的外来差别所增加的误差。

当过了一段时间以后，用它们来进行重复测试时，形象化类比量表的另外一个优点在于，被试很难或者不可能精确地认可他们

过去的反应。比如前讲过的那个关于自信题项的例子，该题项是一个像利克尔特量表这样的多项反应，被试要记住他们所做的在这五个选项中的选择，或许会并不困难。然而，除非被试在形象化类比量表中所选择的是两个端点，否则就很难准确地回想出在一条没有特征的线上所标记的点在哪里。如果研究者注意到被试可能会产生前后一致的反应，这可能会是有利的。或许，在呈现一个实验干预之后，以相同方式所刺激的被试会选择与先前的刺激一样的反应。形象化类比形式基本上排除了这种可能性。如果对于实验被试来讲，后处理反应总是带有一定的偏向性（例如，通常以相同的方向），而对于控制组来讲，前处理反应是随机地偏向性的话，那么选择形象化类比量表就会有助于检测这个被其他方法所疏漏了的微妙的现象。

形象化类比量表被经常用作单一题项测量。这在排除任何内部一致性的检测方面有相当大的缺点。对于一个单一题项的测量，信度只能够通过在第 3 章所描述的测试-重测方法或者通过与已经建立了心理测量特性的同一特征的测量进行对比来测定。前一种方法会遭受前面所讨论的测试-重测评估问题，特别是不可能区别测量过程中的不稳定性和所测量现象的不稳定性。后一种方法实际上是一个结构效度比较。然而，因为信度是效度的一个必要条件，如果效度是显而易见的话，那么我们就可以推测出信度。但是，一个更好的策略可能是编制出复合的形象化类比题项，以便能够测定到内部一致性。

数字化反应形式与基本的神经加工

最近发表在《自然》上的由若瑞、普瑞弗特斯和伍密利塔（Zorzi, Priftis & Umilita, 2002）所进行的一项研究表明，某些反应选择可能与大脑怎样加工数字信息有关。根据这些作者的观点，连续排列的数字，正如在典型的利克尔特量表中那样，不仅以它们的数字值来表示数量而且还以它们的位置来表示数量。他们认为，数字的形象化线条不但是一个方便的表达形式，而且与基本的神经加工相对应。他们观察到，那些有各种大脑损坏从而损伤了视野中的空间知觉的人们，在简单的、形象化呈现的数学问题中会系统地犯错误。视觉上的不正常与所犯的错误类型高度相关。对于无法觉察左边视野的个体，要求当其在一条线上排列的两个值之间

标记一个中点时，他们经常错误地标记"在右边"。例如，当被问到所标记的"3"和"9"这两个点之间的中点是什么时，就会错误地偏向右边（例如，偏向较高的值）。把量表从高到低反过来，还是继续产生向右偏向（现在，偏向较低的值）。当同样的任务以非形象化的形式呈现时——例如，通过询问3和9的平均值是多少——这种模式就没有出现。实际上，当不以视觉方式呈现时，这些个体在数学运算中没有表现出任何缺陷。没有视觉异常的控制组被试并没有表现出那些有大脑损坏的被试所表现出的偏向模式。作者得出结论认为，他们的工作表明了，"心理数字线不仅仅是一个隐喻的有利证据"并且"以空间术语来思考数字（已经由伟大的数学家们所报道了）可能会更有效，因为它是以数字的实际神经表征为基础的"（Zorzi et al., 2002, p. 138）。虽然，这个研究本身不可能保证严格的结论，但是它产生了一个非常令人感兴趣的假设，即评价一条直线上的数字串可能与基本的涉及数量评估的神经机制相对应。如果实际情况就是这样的话，那么作为一排数字而呈现的反应选项可能会有特别的价值。

双极（binary）选项

另外一个普遍的反应形式是，让被试在每个题项的双极选项中做一个选择。虽然具有相等权重的题项也可能有双极反应选项，但是早先的瑟斯顿量表和加特曼量表中的例子使用了双极选项（"赞同"与"反对"）。例如，要求被试核对一个题项清单上他们认为适合他们自己的所有形容词。或者，要求他们对他们在特定情境中所经历的一个情绪反应题项集做出"是"或"否"的回答。在这两种情况中，那些反映了有一个共同的潜在变量的题项的选项（例如，像代表抑郁的"悲伤"、"不幸福"以及"忧郁"这样的形容词），应该被合并为那个结构的一个分数。

双极反应的一个主要缺点是，每一个题项只能有最小的可变性。类似地，每一对题项也只能有共变的两个水平中的一个：赞同或反对。回想一下第3章中，我们讲过，由多个权重相等的题项所组成的一个量表的方差完全等于各个题项的协方差矩阵中的所有元素的总和。对于双极题项，由于在可能的方差和协方差中的限制，每一个题项都对那个总和有着珍贵的贡献。如果题项是双极的，其实际的结果就会是，需要更多的题项来获得相同程度的量表

方差。然而,双极题项通常非常容易回答。因此,对于任何题项来讲,所给予被试的负担就比较轻。例如,大多数人能够很快决定某些形容词是否是对他们自己的合适描述。结果,被试经常愿意完成双极题项,而不愿意完成那些需要集中精力来辨析的题项。因此,两个双极题项形式可能使研究者通过聚合更多题项的信息来获得量表分数中的足够方差。

题项时间结构(item time frame)

关于量表的适当形式的另外一个话题是,具体化的或者隐含的时间结构。在这套书的另外一卷中,克里和麦克格瑞斯(Kelly & McGrath,1988)已经讨论了考虑不同测量的即时特征的重要性。有些量表会涉及时间结构,暗示着一个广泛的时间观点。例如,控制点量表经常包含那些暗示着对于因果关系的一个持久信念的题项。像"如果我采取正确的行动的话,我就能够保持健康"(Wallston et al. , 1978)这样的题项,就假设这种信念是相对稳定的。作为一个对结果的控制的一般而不是特殊的期望,这与控制点理论特征相一致(虽然,在后来的控制点信念的测量中,有一个向更特殊化的转变——例如,DeVellis, Revicki, Lurie, Runyan, & Bristol, 1985)。而另外一些量表则评价相对短暂的现象。例如,抑郁可能会随着时间而变化,并且测量它的量表也必须得承认这一点(Mayer, 1978)。例如,被广泛使用的流行病研究中心抑郁量表(Radloff, 1977)就使用了一个研究范式,该范式要求被试指出在过去的一周他们体验各种情绪状态的频率。一些量表,例如焦虑量表(例如,Spielberger, Gorsuch, & Lushene, 1970),被编制出不同的形式来评价相对短暂的状态或相对持久的特性(Zuckerman, 1983)。研究者应该主动地而不是被动地选择一个量表的时间结构。对于这个过程,有一个理论指导是非常重要的。所研究的现象是个体个性的基本而持久的方面,还是可能依赖于改变中的环境?量表的目的在于检测发生在一个短暂的时间结构中的突然变化(例如,在观看了一场悲剧电影之后增加了消极的影响)还是发生在一生中的变化(例如,随着年龄的增长在政治上逐渐倾向保守)?

总之,题项的形式,包括反应选项和指导,应该反映研究中的

潜在变量的本质以及该量表的使用意图。

步骤 4：让专家评价最初的题项库

到目前为止，我们讨论了三个话题：清楚地表达所要研究的现象的必要性，生成一个合适的题项库的必要性，以及为这些题项选择一个恰当的反应形式的必要性。量表编制过程的下一步将是请一些这个领域的专家来评价题项库。这个评论的目的在于使该量表的内容效度（见第 4 章）最大化。

首先，让专家评价你的题项库，这能够确认你对现象的定义，或者相反，证明你的定义无效。你可以请你的专家组（例如，广泛地从事研究问题中的结构或相关现象工作的同事）来评价，并请他们对每一个题项与你所要测量的东西之间的相关程度做出评估。如果你正在编制一个由各个分量表所组成的量表来测量多个结构的话，这点尤其有用。如果你在编制你的题项之中非常仔细的话，那么专家在确定哪个题项与哪个结构之间相对应时，麻烦似乎就会很少。本质上，你对每一个题项所测量的东西的看法就是你的假设，而专家的反应只是确证数据或者驳斥数据。即使所有题项的目的都在于考察一个单独的特征或者结构，专家的评价也是非常有用的。如果专家对一个题项增加了一些你本没有打算要包括的内容，在完成最终量表时，被试也会这样做。

要获得题项相关的评价，通常要把你对所研究的结构的操作性定义提供给专家组。然后请求他们就这些题项与你所定义的结构之间的相关性对每个题项进行评价。对于每个题项，这就有必要把相关程度评价为高、中、低。此外，你可以邀请专家对他们所认为合适的每个题项进行评价。这使得他们的工作要更难一些，但是可以获得极好的信息。例如，关于为什么有些题项很模糊的一些中肯而有见地的评价，会给你一个新的观点来审视你准备如何来测量这个结构。

评价者也能够评价题项的清晰性与简洁性。一个题项的内容或许与所测量的结构有关，但是其措辞可能有问题。这关系到题

项的信度，因为一个模糊或者要么不清楚的题项，在很大程度上比一个清晰的题项能够反映与潜在变量无关的因素。在你对评价者的指导语中，如果他们愿意的话，请他们指出拙劣的或者混淆的题项并且建议其他的措辞。

你请的评价专家所能提供的第三个帮助在于，指出你还没有考虑到的对这个现象进行研究的方法。或许有一整套方法被你忽略了。例如，在一个关于健康信念的题项库中，你或许已经建立了关于疾病的许多题项，但是没有把受伤考虑为健康的另外一个相关因素。通过对你已经用来测验所研究的现象的各种方法进行评论，评论专家能够帮助你使你的量表的内容效度最大化。

关于专家组评价的最后一句警告：作为量表的编制者，对于专家组的意见，是接受还是抛弃，最终的决定权在于你自己。有时，有关内容方面的专家可能不理解量表构建的规则，这就可能导致坏的建议。从那些没有量表编制经验的同事那里，我经常收到这样的建议，即删除关于同一事物的题项。正如早先所讨论的那样，从一个题项库或者一个最终的量表中去除所有冗余，可能是一个严重的错误，因为冗余是内部一致性的一个整体部分。然而，这个评价或许表示题项的措词、词汇以及句子结构太相似从而应该得以改善。一定要仔细注意你收到的来自有关内容的专家的所有建议。至于怎样使用他们的建议，你自己得做出明智的决定。

在量表编制过程的这一点上，量表的编制者有了一套已经由专家评论过并因此修改过的题项。现在到了进入下一个步骤的时候了。

步骤5：考虑确认题项的包含性

显然，编制量表问卷的中心是题项集合，因为所要编制的量表将从这些题项中形成。在同一个问卷里面包括一些额外的题项，这或许有可能，并且也相对方便，因为这有助于确定最终的量表的效度。这里，至少有两类题项需要考虑。

在量表中，编制者应该选择的第一类题项是用来帮助自己发

现问卷中的瑕疵或者问题。被试可能不会按照你所假设的原由来回答最初研究的题项。可能有其他的动机会影响他们的反应。早点知道这一点是有好处的。其中一种能够被相当容易评价的动机是社会赞许性(social desirability)。如果一个人被强烈地激发按照社会认可的积极方式来呈现她或他自己的话,题项反应可能会被歪曲。包含了社会赞许性的量表就要求研究者去评价每个题项受社会赞许性影响的强烈程度。显著地与所取得的社会赞许性分数相关的题项应该被考虑为排除的对象,除非有充足的理论依据认为这些题项代表其他的东西。斯锥罕和盖贝斯(Strahan & Gerbasi, 1972)已经编制出了一个简单而又有用的社会赞许性量表。这个包含 10 个题项的量表能够被方便地插入一个问卷之中。

还有其他可资利用的题项帮助发现不当的反应倾向(Anastasi, 1968)。明尼苏达多项人格量表(minnesota multiphasic personality inventory),或者 MMPI (Hathaway & Meehl, 1951; Hathaway & McKinley, 1967),就包括几个其目的在于检测各种反应偏见的分量表。在某些情况中,包括这些类型的分量表可能是合适的。

在这一阶段考虑包含性的另外一类题项属于量表的结构效度。正如在第 4 章中所讨论的那样,如果理论认为你所要测量的现象与其他的结构相关,那么该量表的分数与那些其他的结构的测量能够被作为其效度的证明。在这个阶段应该尽可能包括相关的结构,而不是在建立了最终的量表之后再增加一个单独的确认工作。这种合成的关系模式能够为效度要求提供一些支持,此外,又提供了一些线索来帮助理解为什么这套题项不像预期的那样起作用。

步骤 6:在一个试测样本中测试题项

在确定了你的问卷之中已经包含了与结构相关的题项以及效度题项以后,它们必须与新题项库一起在被试中试测。被试的样

本应该大些。多大才算大？关于这一点，很难找到一个一致的意见。先让我们来讨论一下关于一个大样本的基本原理。农纳利(Nunnally,1978)指出，在量表编制中的基本的取样问题，就是从一个假设的范围中选取一些题项作为样本（如，Ghiselli, et al.，1981)。为了强调题项的充分性，样本应该足够地大以便排除作为一个重点关注的问题的被试方差。他建议 300 个人是一个合适的数字。然而，实践经验表明，量表也能成功地用较小的样本来检测。题项的数量以及将要析取的量表的数量也关系到样本的大小问题。如果从一个有大约 20 个题项的题项库中仅仅抽取一个单一的量表，那么 300 个被试以下也就足够了。

使用太少的被试会有几个危险。首先，题项中共变形式可能不稳定。当其被用在一个单独的样本上时，一个表面上增加了内部一致性的题项可能实际上是一个无用物。如果因为其对阿尔法(α)的贡献，题项被选择包含进来（正如经常可能发生的那样），那么一个小的试测性样本会对内部一致性提供一幅不准确的乐观图景。当被试与题项之间的比率相对低，并且样本容量并不大时，题项之间的相关会由于巧合因素而受到相当大的影响。当重新测试在这种情况下编制出来的含有这样的题项的量表时，那些使有些题项最初看起来很好的偶然因素将不再起作用。结果，偶然取得的而不是最初的试测所取得的 α 可能比预期的要低。类似地，一个可能是很好的题项或许会被排除，因为其与其他题项的相关会完全由于偶然因素被而削弱。

小样本的另外一个潜在的缺点是，试测样本可能不代表这个量表打算测量的人群。当然，试测样本很大也可能有这样的情况，但是一个小的样本更有可能排除某些类型的个体。因此，量表编制者应该既要考虑试测样本的大小也要考虑其成分。正如第 3 章中所讨论的那样，一个仔细的研究者应该选择用一个 G-研究来表达一个量表跨民族的普遍性（或者某些其他方面）。

并不是所有样本的非典型性都一样。至少在两个不同的方面，一个样本无法代表更大的人群。第一方面涉及出现在样本与目标群体中的特征的水平。例如，一个样本与预期的人群相比，可能会代表着一个较狭窄的特征范围。这个压缩的范围可能也是不

均匀的，因此对于样本而言，该量表所获得的平均值就会比对于目标人群而言的要高一点或低一点。例如，关于适合喝酒的合法年龄的看法，在一个大学校园中和在一个大的社区中相比，可能变化很大。关于这个特征的不具代表性的平均值并非必然使样本失去量表编制的目的。它可能会产生对于量表均值的不精确的期待，但是其还是为量表所拥有的内部一致性提供了一幅精确的图画。例如，这种类型的样本或许仍然会得出关于哪些题项之间有着非常强的相关的正确结论。

样本不具有代表性的另外一个麻烦是，一个与目标群体有质上而非量上差异的样本，尤其是，其题项或结构中的关系与那些目标群体中的关系不同的样本，有理由值得关注。如果一个样本非常特殊，对于一般的人来说题项就会有一个不同的意义。题项中的相关方式可能反映了由样本成员所共享的不寻常的性质，但是在更大的群体中却很少。换句话说，呈现出内部相关的题项（例如，通过因素分析）的分组可能是非典型的。稍微正式一点来讲，如果一个样本在重要的方面与目标群体不一样，把变量与实际分数关联在一起的潜在的因果结构可能是不同的。考虑一些相当明显的例子：如果所选择的样本的成员不能理解一个重新出现在题项中并且与结构有关的重要单词，那么他们的反应将只能告诉我们很少的信息或无法告诉我们在不同的情境中这个量表如何发挥作用。*sick* 这个单词在美国表示“生病的”但是在英格兰却表示“作呕的”（例如，反胃）。因此，为一个群体编制的关于生病的一套问题，对于另外一群人来说可能有明显不同的意思。如果该量表是关于通常与作呕无关的特定健康问题（例如，风湿病），并且如果样本是英国人的话，那么使用 *ill* 这个单词的题项会聚合在一起，因为其有截然不同的意思。另外一方面，一个美国样本就不可能把关于生病的陈述和其他与健康有关的题项区别开来。即使是在美国，同一个单词也可能有不同的意思。例如，生活在农村的南方人，*bad blood* 有时候被用作对性病的一种委婉说法，而在这个国家的另外一些地方，它的意思是“仇恨”。如果一个讨论“亲戚之间的 *bad blood*”的题项在农村的南方人与其他样本之间分别进行测试的话，其结果就可想而知了。

第二种类型的样本非典型性的结果会严重影响一个量表的编制工作。形成的潜在结构——对于量表信度非常重要的题项中的共变模式——可能是在编制过程中因样本而产生的一个巧合。如果研究者有理由相信,试测样本中的题项的意义可能不是稍大一点群体中的题项意义的典型代表的话,在解释从那个样本中所获得的发现时要非常小心。

步骤 7:求题项的值

在编制好一个初步的题项,仔细地对之进行检查,并且将它在一个适度大小而且有代表性的样本中测试以后,对每个题项进行评价,以便确定合适的题项从而组成一个量表。从很多方面来讲,这是量表编制过程的中心环节。在重要性方面,题项评价或许仅次于题项编制。

初步检查题项的分数

在编制题项时,我们讨论了量表题项的一些理想的性质。让我们再来讨论一下这个话题。在一个题项中,我们所寻求的根本的性质是一个潜在变量的实际分数之间的高度相关。这直接来自于第 3 章中关于信度的讨论。我们无法直接评价实际的分数(如果我们能够的话,我们可能就不需要一个量表了),并且因此不能直接计算它与题项之间的相关。然而,我们可以根据到目前为止已经讨论过的正式的测量模型来进行推测。在第 2 章中讨论平行测试的时候,我讲到任何两个题项之间的相关等于这两个题项之间的一个与实际分数之间的相关的平方。这个平方值就是每个题项的信度。因此,我们可以从题项之间的相关来获得与实际分数之间的相关。题项之间的相关越高,每个题项的信度就越高(它们与实际分数之间的关系就越紧密)。各个题项越可信,由这些题项所组成的量表也越可信(假设它们有一个共同的变量)。因此,在一套量表题项中我们首先要寻求的性质是它们之间的内部高度相关。一种对题项的内部相关进行测定的方法是相关矩阵。

反向记分

如果有题项与其他题项之间的相关是负的，那么应该考虑对这些题项进行合适的反向记分。早先我就提出，以相反方向措辞的题项可能造成问题。然而，有时候我们可能漫不经心地就把负相关的题项处理掉了。例如，如果我们最初期待两个单独的题项组(例如，关于幸福与悲伤)，由于某种原因决定它们应该被合并为一组的话，这种情况就有可能发生。于是我们可能处理掉那些与新合成的结构(例如，情感)同等相关的陈述，但是有些陈述可能是积极的而有些则是消极的。“我很幸福”与“我很悲伤”都适合情感。但是，它们却相反。如果在我们的量表中我们想要好的分数来测量幸福，那么我们就会必须给认可“幸福”的题项以高的分值而给认可“悲伤”的题项以低的分值。做到这一点的一种方法是，使反应选项的文字描述(例如，“强烈反对”、“中等程度地反对”等)对于所有题项来说都始终以相同的顺序出现，而对于与它们相联系的分值则要么按照升序排列要么按照降序排列，这取决于题项，如下所示：

(1)我经常感到悲伤。

6	5	4	3	2	1
强烈地反对	中等程度地反对	温和地反对	温和地赞成	中等程度地赞成	强烈地赞成

(2)大多数时候，我是幸福的。

1	2	3	4	5	6
强烈地反对	中等程度地反对	温和地反对	温和地赞成	中等程度地赞成	强烈地赞成

这个过程可能会使被试感到迷惑。当意识到对于所有的题项来讲它们都是一样的之后，人们就会忽略这些词语。然而，改变一下描述的顺序或许更好一些(例如，对于一些题项，从左到右，从“强烈地反对”到“强烈地赞同”；而对于另外一些题项则反过来)。另外一种方法是，对于所有题项，使文字描述和它们相应的分值都

一样，但是在数据编码的时候给某些题项不同的值。在编码的时候改变某些题项的分数，这项工作既单调乏味又有潜在犯错的倾向。对于每个被试来讲，每一个要被反向记分的题项在反向记分中要给予特别的注意。这就为犯错误留下了大量的机会。

最简单的反向记分方法是，在数据被输入计算机后立即就进行反向记分。一些计算机软件能够处理所有被试数据的所有反向记分。如果反应选项有数字形式的值并且理想的转换是把这些值的顺序反过来的话，那么就可以使用一个简单的公式。例如，假设其模式为使用利克尔特量表的一套关于情绪的题项，从 1 到 7 记分，大的数字表示赞同。进一步假设，为了便于理解，积极的情绪题项和消极的情绪题项都使用了这种相同的反应模式。然而，如果认可积极的情绪与高的分数相联系的话，那么这个量表本质上就是一个积极的情绪量表。认可一个积极的情绪会得到一个高的值，而认可一个消极的情绪会得到一个低的值。如果对于所有消极情绪题项来说，7 这个反应值被换成 1，6 换成 2，如此等等，其就是可能得到的结果。这种转换可以用以下这个公式通过从已有的分数中创建一个新的分数来完成：

$$新分数=(J+1)-已有分数$$

这里新分数和已有分数分别指转换后的和最初的分数，而 J 是最初的反应选项的数目。在所举的例子中，J 就等于 7 而 $(J+1)$ 等于 8，从 8 中减去分数 7 的结果是 1，减去 6 的结果是 2，如此等等。

题项中的有些负相关可能无法通过反向记分来修正。例如，对一个特定的题项进行反向记分会消除一些负相关，但是会产生另外一些负相关。这通常表示有些题项不是简单地属于量表，因为它们不是一贯地与另外一些题项相关。如果没有反向记分这种模式来消除负相关的话，那么在一个相似的集合中，与一些题项正相关而与另外一些题项负相关的任何一个题项都应当被排除。

题项-量表相关

如果我们想得到一套内部高度相关的题项的话，那么每一个

题项都应该稳定地与其余的题项集合相关。对于每一个题项，我们可以通过计算其题项-量表相关来检查这种性质。有两种类型的题项-量表相关。修正的题项-量表相关使被计算的题项与除自身以外的所有量表题项相关，而未修正过的题项-量表相关使问题中的题项与整个候选题项集合相关，包括它自身。如果一个量表中有 10 个题项需要考虑的话，对于这 10 个题项中的任何一个，修正的题项-量表相关将由其与其他 9 个之间的相关来组成。未修正的题项-量表相关将由其与所有 10 个之间的相关来组成。理论上，未修正的值告诉我们该题项在整个量表中的代表性如何。例如，这与求一套 IQ 测试的分测试与整个测试之间的相关从而决定这个子测试是否是具有代表性相类似。然而，虽然一个未修正的题项-量表相关提供了好的概念上的意义，但是实际的情况却是，题项包含在"量表"中会使相关系数膨胀。一套量表中题项的数量越少，详细调查中的题项的包含性或排除性所带来的方差就越大。总之，检查修正的题项-量表相关或许是明智的。对于这种相关，一个分值高的题项比一个分值低的题项更理想。

题项方差

量表题项的另外一个有价值的特征是相对高的方差。列举一种极端的情况，如果所有的被试对一个特定的题项的回答完全一样，那么在个体中就根本无法辨别所测量的结构的不同水平，并且其方差将会是 0。相反，如果试测样本在所研究的特征方面是不同的，那么一个题项所获得的分数的范围也会是不同的。这就暗示着一个相当大的方差。当然，通过加入误差成分来增加方差，这种做法并不是我们想要的。

题项的平均值

在选项的分数范围内，接近中心位置的平均值也是理想的。例如：如果对每一个题项的反应选项的值的范围是从 1 到 7，选 1 表示强烈的反对，而 7 表示非常同意，那么一个接近 4 的平均值是很理想的。如果一个选项在这个范围的极端值附近，那么这个选项可能无法检测这个结构的某些值。例如，如果很多人都选择 7 这个分数值，则暗示着这个选项的措辞不够有力（例如，很难找到

不同意该题项的人)。

通常,与反应范围的极端值靠得太近的平均值题项将有低的方差,并且那些在一个很小的范围内变化的题项与其他题项的相关是很差的。正如前面所讲的那样,一个没有变化的题项不能共变。因此,任何不平衡的平均值或者低的方差,都将会减少一个题项与其他题项之间的相关。可以把注意力主要放在题项之中,作为对它们的潜在值测量的一个关系模式。然而,一旦根据相关而对题项做出了一个试探性选择的话,那么检查平均值和方差将是一个有用的双重检查(double-check)。

因素分析

一套题项集合并不必然是一个量表。题项可能没有共同的潜在变量(如在一个索引或突然出现的变量中)或者可能有好几个。确定隐含在一个题项集之中的潜在变量是非常关键的。例如,作为阿尔法的一个基础的假设是,这个题项集合是单维度的。确定哪些题项集(如果有的话)组成一个单维度的集合的最好办法是因素分析。对整个章节(见第 6 章)的有益性来说,这个话题是足够重要的。因素分析要求大量的样本,量表的编制总体上来讲也是需要的。如果用于因素分析的被试太少的话,整个量表的编制过程可能就会受到影响。因此,某些类型的因素分析通常是这个时期的量表编制的一部分。

阿尔法系数(coefficient alpha)

一个量表最重要的指标之一是信度系数,阿尔法。实际上,迄今为止所讨论的所有各个题项问题——非中心平均数,不好的差异性、题项之中的负相关、低题项-量表相关、以及弱的内部题项相关——都将会减少阿尔法。因此,在我们选择了题项,即去除了差的题项而保留好的题项之后,阿尔法是对我们所做的工作成功与否的一种评价方法。阿尔法表示由实际分数引起的在量表分数中的方差的比例。用于计算阿尔法的方法有好几种,但在自动化的程度方面却不同。一些计算机程序含有计算阿尔法的题项分析程序。在 SPSS 中,有信度程序来计算整个量表的阿尔法以及所有 k

−1 个版本的阿尔法(例如,去掉一个题项之后的每一个可能的版本)。这个程序也提供修正过的和未修正过的题项-量表相关。作为相关程序的一个特征,Proc Corr,SAS 包括阿尔法计算。在 Proc Corr 中,通过选择阿尔法选项,在伴随变量(例如,说明)的陈述中所列举的变量将被作为一个量表,并且将会为整个题项集合以及所有可能的 $k-1$ 个题项集合计算阿尔法。同时也会得到题项-量表相关。

计算阿尔法的另外一种方法是用手工计算。如果单个题项以及作为一个整体的整个量表的方差是有用的话,那么它们就能被带入在第 3 章中所讨论的第一个公式中去求得阿尔法。或者我们可以使用斯皮尔曼-布朗公式,这个公式也是在第 3 章中介绍的。这个公式使用来自相关矩阵中的可用信息而不用来自方差中的信息计算阿尔法。这个方法的一个缺点是,相关是标准化的协方差,并且对每个题项进行标准化可能会影响到阿尔法的值。如果我们严格遵守平行测试的模型,那么这就是不合逻辑的,因为相关是被假设为相同的。然而,实际上它们永远不会精确地相等。基本 Tau 相等测试模型没有要求题项中相等的相关,只是要求协方差相等。因此,由于误差而引起的每个题项的方差的比例在这个模型下就不会变化。然而,由于斯皮尔曼-布朗公式的操作原理实际上是平均的内部题项相关,并且 Tau 相等测试模型所隐含的一个条件是,平均的题项—量表相关对每个题项都是相等的,所以仍然没有问题。尽管如此,但是在以协方差为基础和以相关为基础的计算模型所取得的阿尔法值之间可能存在着小的(或许有时是大的)差别。因为协方差矩阵使用一个比较单纯形式的数据(没有标准化),因此它更令人喜欢并且经常被使用。

理论上,阿尔法的取值可以从 0.0 到 1.0,但是其不太可能取得这两个极端值中的任何一个。如果阿尔法值是负的,那就意味着出现了什么错误。一个可能的问题是在题项中的负的相关(或者协方差)。如果这种情况发生了,正如在本章早先所讲到的那样,尝试反向记分或者删除一些题项。农纳利(Nunnally,1978)建议把 0.70 这个值作为阿尔法的比较低的可接受边界值。在已经出版的量表中,我们会经常看到有比较低的阿尔法的量表。不同

的方法学家和研究者开始寻求不同水平的阿尔法值。对于研究量表，我个人觉得合适的范围如下：低于0.60，不能接受；0.60～0.65，不理想；0.65～0.70，最低程度的可接受；0.70～0.80，可观的；0.80～0.90，非常好；大量地超过了0.90，我们应该考虑缩短这个量表(见下一节关于量表的长度)。我要强调的是，对阿尔法值的这种分段，是我个人的经验并带有很大主观性。我们无法为它们找到严格的理论基础。然而，它们反映了我的经验并且似乎与其他研究者的评价有大量的交迭。我所建议的值适合稳定的阿尔法。在编制过程中，根据其对阿尔法的贡献，直接地或者间接地选择题项。题项中有些明显的共变可能是由于偶然因素造成的。因此，在编制阶段尽量寻求那些比你所期待的阿尔法要高一点的阿尔法是明智的。那么，当其被用在一个新的情境中的时候，如果阿尔法被恶化了，它们也仍然可以接受。正如早先所讲到的那样，如果发展性样本太小，研究者应该尤其要注意，在量表编制阶段所获得的最初的阿尔法估计值可能不稳定。正如我们将要看到的那样，当组成量表的题项数量太小时也会出现这种情况。

当人们在编制一个要求严格的精确的量表时，我所建议的阿尔法的"合适范围"就不适用了。临床情境就是一个例子。我所建议的阿尔法范围适合与群组数据一起使用的研究工具。例如，其阿尔法为0.85的量表可能完全适合在与所测量的结构有关的对比组研究中使用。单个体的评价，尤其是根据这个评价做出重要的决定的时候，则要求更高一些的标准。例如，用于个体诊断、雇佣、学术地位或者其他重要的目的的量表，或许要有相当高的信度，应在0.90～1.0的范围内。

在某些情况下，例如当量表由一个单题项组成的时候，把阿尔法用作信度的指标就不可能。如果可能，应该做一些信度评价。测试—重测相关可能是这种单一题项事例中的惟一选择。虽然这个信度指标并不完美，正如第3章中讨论的那样，但是明显要比完全没有信度评价要好。如果可能，一个更好的方法是用不止一个题项来构建这个量表。

步骤 8：优化量表长度

量表长度对信度的影响

在量表编制的这个阶段，研究者已经有了一个可接受信度的题项库。一个量表的阿尔法受两个特征的影响：题项中的共变的程度以及量表中的题项数量。对于那些有题项-量表相关的题项，如果其与平均的内部题项相关大约相等(例如，非常有代表性的题项)，那么增加更多的题项将增加阿尔法，而减少题项将降低阿尔法。一般而言，短一点的量表比较好，因为它们给被试的负担较少。另一方面，较长的量表更可靠。显然，增大这两个方面的一个就会减少另外一个。因此，量表的编制者应该考虑一下简短性和信度之间的最佳平衡。

如果一个量表的信度太低，那么简短性就没有意义。实际上，被试或许会更愿意回答一个有 3 个题项的量表而不是一个有 10 个题项的量表。然而，如果研究者不能给从较短的量表中获得的分数以任何意义的话，那么量表就失去了价值。因此，只有当研究者有“多余的信度”时，才能对量表长度进行缩减。实际情况则是，人们更注意以更低的成本编制信度稍许低一些的短量表。

删除“差的”题项的影响

删除“差”的题项实际上是增加了阿尔法还是稍微降低了阿尔法，这取决于被删除的题项有多差，以及这个量表中的题项数量。考虑一下较多的题项或较少的题项的影响，这些题项都是同样“好的”题项——即与它们的对应物有可比较的相关：具有较少题目的量表，阿尔法会随题目的增加或减少产生大幅度的改变。如果 4 个题项中的平均内部题项相关是 0.50 的话，那么阿尔法就等于 0.80。如果只有 3 个题项的平均内部题项相关为 0.50 的话，阿尔法就会下降到 0.75。有同样的平均内部题项相关的 5 个题项，阿尔法就为 0.83。对于平均内部题项相关为 0.50 的有 9 个、10 个和

11 个题项的量表，阿尔法就会分别为 0.90、0.91 和 0.92。在后面几种情况中，阿尔法不仅要高一点，而且它们的值也靠得更近。

如果一个题项与其他题项之间有一个低于平均值的相关，去掉它会增高阿尔法。如果其与另外题项之间的平均相关只比总体平均值低一点点(或者相等，或者高一点)，那么保持这个题项将增加阿尔法。我在前面就说过，一个有 4 个题项的量表会取得一个 0.80 的阿尔法，平均内部题项相关为 0.50。对于那个将被删除的题项来讲，与另外 3 个题项之间的平均相关为多低才能有助于阿尔法而不是妨碍阿尔法？首先，考虑一下以下情况，对于一个有 3 个题项的量表来说要获得一个 0.80 的阿尔法，其平均内部题项相关必须为多少？它应该需要为 0.57。因此，在删除了 4 个题项中的最差的一个之后，为了保持阿尔法的值为 0.80，剩下的 3 个题项就需要一个 0.57 的平均内部题项相关。其平均内部题项相关低于0.57的 3 个题项会比其内部题项平均值为 0.50 的 4 个题项有一个较低的阿尔法。假设 4 个题项中的 3 个最好的题项之中有一个 0.57的平均相关，剩下的(因此也是最差的)一个题项与其他 3 个之间的平均相关就必须低于 0.43，这样它的删除就会实际上增加阿尔法(有 3 个内部题项相关的平均值为 0.57 的题项和 1 个内部题项相关的平均值为 0.43 的题项，就会得到这 4 个之中的总体平均内部题项相关值为 0.50)。对于任何大于 0.43 的值，保留第 4 个题项要比降低平均内部题项相关要好。因此，这一个“差的”题项要远比另外 3 个题项差(0.57－0.43＝0.14)才值得删除掉。

现在，考虑一下当一个有 10 个题项的量表并且其阿尔法为 0.80时的情境。首先，平均内部题项相关仅仅需要为大约 0.29，这表明了这样一种方式，即更多的题项抵消了它们之中的较弱的相关。对于一个有 9 个题项的量表来说要获得同样的阿尔法，平均内部题项相关就需要大约为 0.31。为了把其包括进来作为第 10 个题项从而把总的平均内部题项相关降低为 0.29，一个“差的”题项就需要与其余 9 个题项之间有一个大约为 0.20 的平均内部题项相关。如果无法使平均值低于这个值，就会导致题项的增加有利于阿尔法。在这个例子中，9 个“好的”题项和 1 个“差的”题项之间的平均内部题项相关的差为 0.31－0.20＝0.11，比在有 4 个题项

的例子中所发现的差别要小。

量表长度的完善

在实际中，人们如何完善量表的长度呢？显然，应该首先考虑删除对内部一致性贡献最少的题项。这些可以通过很多种方法来确定。SPSS信度程序以及在SAS中的Proc Corr的阿尔法选项可以显示删除一个题项对整个阿尔法造成的影响。删除后对阿尔法产生最少的消极影响或者最大的积极影响的那些题项，通常是最好的首选删除题项。题项-量表相关也可作为确定哪些题项应该被舍去的一个标准。那些有最低的题项-量表相关的题项应该首先被删除。SPSS也为每个题项提供了一个平方多重相关(squared multiple correlation)，这是通过求一个题项与所有剩余题项的回归而获得的。这是对题项的集体性的评估，即这个题项与其他题项共享方差的程度。就如题项-量表相关的情况一样，具有最低的平方多重相关的题项是需要删除的主要对象。一般而言，题项质量的这些不同指标是集中在一起的。当这个题项被删除以后，一个差的题项-量表相关一般会伴随着一个低的平方多重相关以及阿尔法值的较少的降低，或者甚至是一个增长。量表长度影响阿尔法的精确度。在实践中，所计算出来的阿尔法是依据测量假设与实际数据之间的合适性而对信度进行的一个估计。我们已经注意到，当包括更多的题项时，阿尔法会增长(除非它们是相对差的题项)。此外，作为对信度的一个估计的阿尔法的信度，也会随着题项数量增长。这就意味着，依据一个较长的量表所计算的阿尔法比一个较短的量表计算出阿尔法更可信。当进行被试间施测时，一个较长的量表比一个较短的量表会有更相似的阿尔法值。在编制过程中，在决定一个量表需要编制的长度时，应该考虑这一事实。

最后，在试图优化量表的长度时应该为阿尔法留一个安全余地，记住这一点是重要的。当用这个量表对一个样本而不是对其最终的被试进行测量时，阿尔法可能会减少一些。

样本分离(split sample)

如果试测样本足够大,就有可能把它分离为两个子样本。一个可以作为最初的试测样本,而另外一个可以被用来反复核对结果。因此,从第一个子样本中所获得的数据可以用来计算阿尔法,评价题项,完善量表长度,以及编制出看起来最理想的量表的最终版本。第二个子样本可以用来重复这些发现。对要保留的题项的选择无须以第二个子样本为基础。因此,这组计算出来的阿尔法以及其他数据就不会出现早先所讨论的偶然性效应,例如阿尔法膨胀。如果这两个组之间的阿尔法保持相当稳定,假设这些值没有被偶然因素所歪曲的话,你就可以更放心。当然,这两个子样本可能要比两个完全不同的样本要相似得多。从整个试测样本中分离出来的子样本,可能会代表相同的人群;相反,一个完全新的样本或许代表一个略微不同的人群。同样,这两个子样本的数据收集时期也不会由于时间而分离,而一个试测样本与一个完全不同的样本几乎总是会分开。另外,适合一个子样本的数据收集的特定情境也相同地适合另外一个。这些特定的情境的例子包括特定的研究人员、物理环境以及问卷打印的清晰性。同样,这两个子样本可能是完成两套题项的两组被试,要完成所有量表题项包括来自最初的题项库中最后要舍弃的题项。如果舍弃的题项对量表题项有影响的话,这些影响就可以在两个样本之间进行比较。

尽管这两个合成的子样本有独特的相似性,但是分离试测样本来重复已有的发现,提供了关于量表稳定性的有用信息。这两个子样本在一个重要的方面有区别:作为题项选择的数据基础的第一个子样本,存在着题项中可靠的共变有可能与不稳定、偶然的因素相混淆。在第二组中,根本就不可能系统地使偶然因素影响信度,因为其数据没有影响题项选择。这一根本差别成为分离样本从而获得有价值信息的最充分的理由。

分离一个足够大的样本的最明显的方法是把它分半。然而,如果这个样本太小而不能分离出两个足够大的均等部分的话,你可以不均等地分离它。较大的那个子样本可以用于更至关重要的题项评价过程与量表构建,而较小的那个用于重复确证。

练　习

假设你正在编制一个关于对蛇的恐怖(fear-of-snake)的量表，采用6个选项的利克尔特反应模式，使用300个被试。虽然对于实际的量表编制来说，会需要更多的题项，但是请做这些练习：

1)创建有10个利克尔特模式题项的一个题项库。

2)对于每一个你所写的题项，估计一下由“持折衷态度的人”(例如，既不是一个蛇恐怖患者也不是耍蛇的魔术师)所认可的利克尔特值是什么？

3)从数据库中选出一个你怀疑会对一个持折衷态度的人造成一个极端反应的题项，并且重新编写它使其引发一个较适中的反应。

4)另外创建10个利克尔特题项来测量蛇恐怖以外的结构。随机把这些题项与原来的10个题项混合在一起，并邀请你的朋友来判断他们认为每个题项是打算测量什么的。

5)使用蛇恐怖或者你的第二个题项库所隐含的结构，直接列举可以观察到的那些能够被用来确认一个测量该结构的量表的行为，并解释你如何使用行为数据来进行确认。

6)如果你的10个蛇恐怖题项有一个0.30的平均内部题项相关，这个量表的阿尔法会是什么？*

7)你怎么运用样本分离来估计和再确认这个量表的阿尔法系数。

* 练习6，阿尔法＝(10×0.30)/[1＋(9×0.30)]＝0.81。

6

因素分析

Factor Analysis

因素分析概述

因素分析的概念描述

因素的解释

主成分与共同因子

验证性因素分析

量表编制中因素分析的使用

样本大小

结　论

在第 2 章讨论能描述量表题项与潜在变量之间关系的不同理论模型时，我提到了一般因素模型。该模型并不假定只有一个潜在变量是题项间所有共变量的来源。事实上，这个模型允许多个潜在变量作为题项集变化的原因。

为了阐明多个潜在变量怎样支撑题项集，我将描述一种具体的、基本上是假想的情境。社会和行为学家感兴趣的许多结构能在多个具体化水平上进行可操作化处理。心理调节（psychological adjustment），情感（affect），消极情感（negative affect），焦虑（anxiety），以及考试焦虑（test anxiety）等术语都是心理结构等级化现象的例子。每个术语都包含量表中的那些题项，并且可能在任何具体化水平上编制量表。可以假定，不同措辞的、具有不同时间结构和反应选项的题项能形成从具体、中等到一般水平的量表连续体。量表编制者期望他们能够选择与预期的变量的特定水平相一致的题项措辞。然后，用因素分析来评价选择过程是否成功。

为了使例子更具体，假设一下有一个包含 25 个题项的、适合用于测量情绪的题项集合。我们关心的是：这些题项是否能够组成一个总体的量表或者许多更具体的分量表。能将所有的 25 个题项组合在一起吗？或者，这些题项是否更适合于分成几个量表来测试不同的情绪状态，例如用来分别测试抑郁、欣快、敌意、焦虑等等？可能这些题项更适合分成几个分别用来测量积极情绪和消极情绪的分量表（例如，从"高兴"到"悲伤"的维度来测试沮丧，或者从"紧张"到"平静"的维度来测试焦虑）。我们怎么知道用哪种方式来处理手边的题项最合适呢？事实上，这个问题的实质是，关于几个情绪状态问题的题项集到底是由一个还是由几个潜在变量支撑的？

在试图回答这些问题时，如果只使用前几章所讨论的方法而不是因素分析方法的话，其结果将会失败。我们将计算有关情绪的全部题项的 α 系数。α 系数能告诉我们一组题项的共同变异是多少。如果 α 系数低，我们就可以找出那些相互之间具有很强的相关的题项，从而形成一个题项子集。例如，我们可能怀疑表示积极情感的题项与表示消极情感的题项之间没有联系，把它们组合在一起会降低 α 系数。更同质的题项子集之间（所有的表示积极

情感的题项或所有表示消极情感的题项)的 α 系数将较高。当然，在某种程度上，我们可能也担心，越具体和越同质的量表相互之间的联系之所以越强，是因为它们仅仅是同一情绪状态的不同方面，这表明这些题项属于同一个量表而不属于相互分离的子量表。

强调一个相关性高的 α 系数并不等于赞同所有的题项都受单一潜在变量的影响。如果一个量表包含 25 个题项，12 个题项反映了一个潜在变量，那么剩下的 13 个题项或许可能反映了另一个潜在变量。在所有题项的相关矩阵中，有些题项之间具有较高相关，有些题项之间的相关则较低。受相同潜在变量影响的两个题项之间的相关会较高，而那些主要受不同潜在变量影响的变量之间的相关则较低。在一个由 25 个题项组成的量表中，不同题项之间的平均相关可能高到足以得出相当大的 α 系数。例如，不同题项之间仅 0.14 的平均相关就足以得出 0.8 的 α 系数。

本章的主题是因素分析。因素分析是一个有用的分析工具，它能告诉我们一些重要的量表特征，这是信度系数所不能告诉我们的；它能帮助我们从实验上决定在这些题项中有多少个结构、潜在变量或因素。

因素分析概述

因素分析的功用较多。它的一个主要功能，像刚刚所说的那样，是帮助研究者决定一组题项中究竟含有多少个潜在变量。例如，以 25 个关于情绪的题项为例，因素分析能帮助研究者决定题项集中表现的究竟是一个更一般化的结构还是几个更具体的结构。因素分析也能通过采用较少的新确立的变量来解释较多的原始变量之间的变异。这意味着精减信息，使得采用少数的几个变量就可以解释变异。例如，一般需要 25 分来描述被试如何回答题项，但是在合并题项的基础上，有可能计算更少的分数(或许甚至是 1 分)来回答题项。因素分析的第三个目的是确定因素的实质性内容或意义(例如，潜变量)，从而能对大量题项集之中的变异进行解释。这种解释可以通过确认相互作用的变量集以及明确定义

潜在变量的潜在意义而得以实现。例如，如果在分析25个关于情绪的题项时出现了2个因素，那么，构成那些因素集的单个题项就能够提供因素所描述的潜在潜变量的线索。

接下来的部分简单地介绍了因素分析的概念。那些想对因素分析有更详细了解的读者可以参考其他作者的文章，如库尔顿(Cureton，1983)、哥萨奇(Gorsuch，1983)、哈曼(Harman，1976)或麦克唐纳德(Mcdonald，1984)。

与因素分析的概念方法类似的例子

为了给什么是因素分析一个直观的概念，我们考虑了我们可能更熟悉的两个例子，这两个例子虽然不那么正式，但却具有大致相似的程序。这个程序的第一个例子有时会在人力资源管理中发生。在人力资源管理中，有时团队内的成员或合作者关心的是这样的问题，即那些貌似不同的各种具体问题背后所存在的共同问题。这时，这个共同的问题需要被识别出来。

例 1

假定一个小的新公司想确定职员认为同事的什么品质最重要。他们认为识别和奖励共同的价值品质对于形成一个协调和合作的工作环境是十分重要的。公司聘请了一个人力资源专家来帮助他们。这个人，我们叫他吉姆(Jim)，召集了公司的10个职员并解释道：他希望他们考虑一下，在所有可能与同事在一起的情境中，包括一起制订提案和报告、一起与潜在客户打交道到以及在咖啡厅共同饮咖啡等等，他们认为同事的哪些品质是最重要的。吉姆建议，在程序开始时，每个职员分别在各自的纸上尽可能多地写下他们认为重要的品质。

在职员写下他们的想法后几分钟，吉姆要求一个志愿者向大家读出他或她的想法。爱丽斯(Alice)说她写下的一个品质是“愿意分享观点”。吉姆感谢她并要求她把写有这个想法的纸贴在墙上。另外一个职员比尔(Bill)，读了一个他认为重要的品质：“幽默感”。这也被贴在墙上。这个程序持续至每个职员都表明了他们认为的合作者应当具有的重要品质为止。这样，人们逐个说出了各种各样的、他们认为同事应具有的重要品质。

当这些工作做完之后，他们把列有每种品质名称的清单贴在墙上。清单列举了如下品质：

愿意分享观点	友好
有幽默感	能被依靠
在工作中永远能选择正确的方式	注意细节
聪明	坚强
不草率	好交际
勤奋	认识大量潜在客户
工作能随时完成	可信赖
有逻辑的思考	有个性
必要时能经受住困难	受过良好教育
为工作做准备	值得信任
给顾客留下良好印象	知道如何穿着
并不争取获得所有荣誉	会讲故事
有趣	天才
有漂亮的车	守诺的人
在这种类型的工作中有大量的经验	愿意为了完成工作需要而长时间工作

这个过程持续了一段时间，很快墙上贴满了 30 多张纸，每一张纸上写着一个职员认为重要的品质。下一步，吉姆问他们是否能够合并一些品质。凯塞瑞(Katherine)指出“聪明”和“天才”是相同的。吉姆拿下清单，将“天才”移到“聪明”的下方。弗兰克(Frank)主张“受过良好教育”也可以分到这一组。几个其他品质也被加到相同的陈述组中。卡拉(Carla)认为“友好”和“给顾客留下良好印象”是相同的但又不同于前一组提到的品质，她主张这两个品质能合并形成一个新的组别。那么，“有趣”也可以加入这个组。“不草率”和“知道如何穿着”构成了第三组。但一个职员认为，把“不草率”和“为工作做准备”放在一组要比把“不草率”和“知道如何穿着”放在一起更恰当一些。这个过程持续到吉姆和职员们得到了几个品质集合为止。事实上，每种被描述的品质都被放

人某个品质集合之中。

然后，吉姆要求职员用一个词或简短的描述性短语给每个品质集合命名。不同品质集合被命名为“智力”、“容貌”、“尽责”、“人格”、“可依赖性”等等。可以假定，每个品质集合代表了一个核心概念，这些概念与职员对另一个人的品质的看法相关联。

例 2

几年后，这个公司准备重复这一做法。经理怀疑事情发生了很大的改变，最初确定的品质集合现在看来可能已不合适。但是，公司又没有像吉姆这样的人力资源管理者。这时，卡若(Carol)，公司的一个执行官，认为想要得到相同信息的一个相对简单的方法可能是编制一个与人们先前进行的测验相类似的问卷。这时，要求职员用“一点也不”、“某种程度上”、“十分”这样的短语来描述他们认为的每个品质的重要性程度。回答问卷调查的职员，现在大约有 150 人。当卡若收回问卷时，她浏览这些问卷并寻找最重要的品质。虽然她发现不同的人认为不同的品质有不同的重要性程度，但还是有某些品质具有相同的重要性等级趋势。例如，如果人们认为“重视细节”是重要的，那么也可能同时认为“为工作做准备”是重要的。那些认为其中一个品质不重要的人，一般也会认为另一个不重要。卡若想弄清楚产生这种现象的原因。她记得，在几年前最初的将纸片贴在墙上的做法中，好像最初产生的组别要比实际所需要的多。她想，一些人的看法可能是相当没有价值的，有时好像同一个人写了多个无价值的品质，这将导致一个整体无价值的类别。她想知道，是否存在一种能够决定可以从职员对其同事的看法的大部分信息中抽取出多少个类别的方法。作为解决这一问题的一种方法，可以像两个她已注意到的选项一样，将职员间相似的看法加以合并。事实上，她在组合相似题项时不仅需要考虑职员观点的内容，也要考虑职员对这些特征题项的评价的相似性。这需要花费大量的时间，而且卡若也不能真正确定她是否选择出了所有的重要的品质集合，但是她能够用问卷这种方式收集一些有趣的想法。

这些方法的缺点

这两个例子虽然与因素分析在概念上存在某些相似之处，但

也存在某些重要的差别。在这两个例子中,测验得到的是一系列已对大量信息重新组织而形成的、易于管理的、更抽象的、但充满意义的类别。因此,每一种重新分类导致的结果是,最初许多个体提出的大量的观点被合并成了几个少数的想法。当然,这两种方法也具有十分明显的缺点。在第一个例子中,研究者没有控制不同个体所产生的看法的性质。例如,虽然性格外向的人并不总是具有洞察力,但通常性格外向的人可能会比性格不那么外向的人提出更多的看法。由于这样或那样的原因,这个过程通常导致一些模糊的、不相关的,或是十分可笑的看法。由于该活动的内容具有的动态性,想要排除一些看法而又期望这么做不会冒犯提出这些看法的人,可能就有点困难。于是研究者可能会让这些看法和那些好的看法一样具有信度。即使这些题项之间具有广泛的联系,一些题项之间也会比其他的题项之间具有更密切的关系。可是,在研究中,所有的题项都或多或少地趋于被平等地对待。如果出现了几个简单的题项,它们可能仅仅基于相似性而组成一个类别。组别可能会被区分出优先顺序,但这通常在所有的参与者一致同意的情况下才会这么做,并且这种优先顺序也依赖于是谁提出这个特征的,而且也可能勉强认为某些类别是不重要的。进一步说,我的关于此类测验的经验表明,参与者存在将每种看法列入某种类别的强烈趋势。几个简洁的类别和一两种孤立的看法并列在一起,使人们觉得其缺少封闭性,所以,在通常情况下,即使没有证据表明剩余的孤立的看法之间存在联系,但参与者还是会将这些孤立的看法组成一组。最后,虽然一个类别可以由类别的具体例子来定义,但有些例子能良好地表征类别,有些例子则不能。

第二个例子避免了上述的一些缺点。卡若能删除一些她认为不相关的题项,虽然这种处理在很大程度上取决于她的主观判断。但至少确认题项的过程在某种程度上是民主的。每个人能在不冒着疏远同事的危险的情况下对每个题项做出评价。分组所依据的不是观点的表面相似性,而是人们是否以一种共同的方式来反映相似的题项集。也就是说,相似性是题项的一个特征(某些题项隐含了相同的看法),而不是被试的特征(对不同的题项做不同反应的人)。参与者认为在一个组别中的某个题项不重要,表明其很可能认为同一组别

中的其他题项也是不重要的。但不同的职员也可能都认为某些题项是重要的。关键的问题是，不管个体如何评价题项的重要性，组内的观点是趋于一致的。事实上，这是卡若建立品质组的基础。对于50份问卷来讲，靠视觉检查来做到这一点，是相当令人气馁的，而且很可能卡若的分类系统也不是所有可能的方法中最有效的方法。对题项来说，具有多大程度的一致性才能被认为是一组？一个职员对同一个潜在品质集合的两个题项做出不同评定的情形（例如，对重要性的赞同与反对），卡若能容忍几次？

因素分析的概念描述

因素分析是一个与上述方法相类似的分类程序，但它是以一系列更加结构化的编制来完成的，并且为数据分析者做出评定提供了更加明确的信息。像刚刚描述的方法一样，因素分析确定了由相似的特征构成的类别。因素分析者的首要任务就是决定需要多少个类别来捕获来自原始的陈述集合中的大量信息。

抽取因子

事实上，因素分析的第一步是假定一个大的类别包含了所有必须的题项（例如，一个概念或类别足以解释反应的模式）；然后评估一个单一概念能在多大程度上解释题项之间的联系；最后，因素分析确认这一单一概念的假设是否恰当。如果一个概念或类别明显不能充分解释题项间的共变，因素分析就会拒绝最初的假设，然后会再确定第二个概念（例如，潜在变量或因子）来解释题项间残余的共变。这一过程要持续至因素不能解释的共变量小到能够接受的程度为止。

第一个因子

怎样完成因子的抽取？这个过程从所有题项间的相关矩阵开始。用这个矩阵作为起始点，因素分析要检验题项间的相关所表示的共变模式。接下来是概念的描述。为了方便阐述，我们略去了一些数学上的细节，所以不能刻板地认为这是计算机进行因素

分析的原理。

正如先前所阐述的那样，这个过程包含了一个最初的假设，即认为单一概念足以解释题项间的相关模式，这等于是一个临时的假设：如果一个模型只有一个单一的潜在变量（例如，一个单因子），并且该潜在变量到每一题项只有一条单独的路径，那么它就能准确地体现因果关系。这进一步表明，这样的一个模型能解释题项间的相关。为了检验这个假设，因素分析程序必须确定每个题项与代表单一潜在变量的因子间的相关，然后再看观察到的题项间的相关能否通过适当的增加因素与每对变量的连接路径来重构。但这个程序怎样能计算到可观察到的题项反应与表示无法直接观察或测量的潜在变量的因子之间的相关呢？

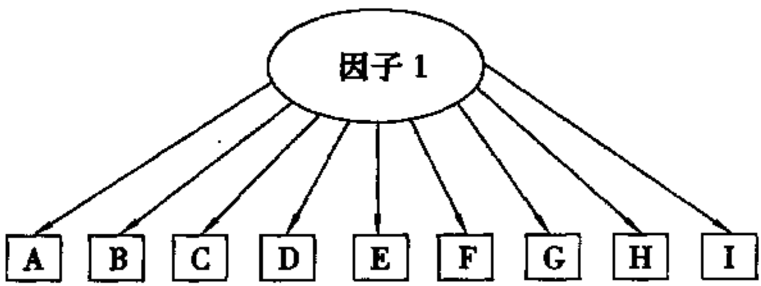


图 6.1　单因子模型

解决这一问题的一种方法就是所有题项反应的总和是这一个包含一切的潜在变量的合理的数量估计值，该潜在变量被假定为可以解释题项间相关。事实上，这一总和是对潜在变量“分数”的一个估计值。因为所有题项的实际分数被认为由一个潜在变量决定，所以来源于所有题项的数量上的联合信息（例如，全部的总和）是对潜在变量的数量值的一个合理估计值。将个别题项分数加在一起得到一个总分，并计算每个题项和所有题项的总和之间的总体题项相关（item-total correlations）是十分简便的。题项总体相关表示的是作为可观察的题项和不可观察的潜在变量（例如，从潜在变量到个别题项的因果路径）之间的相关。随着数值被分配到因果路径上，就能计算出基于单因素模型反映出的题项间的相关。

如果只存在一个潜在变量这一前提是正确的，那么这些由模型得出来的相关应该就是对实际的内部题项间相关的映射。我们可以通过比较映射的相关和实际的相关来评估这一前提的合理性。这等于从以原始数据为基础的相应的实际相关中减去每个映射相关。如果实际的相关与映射的相关存在着差异，那么这表明单因素模型是不充分的，题项间仍存在一些不能解释的共变。

考虑一下一个单一题项对的这样一个序列 A 和 B，它们是一个大题项集的一部分。首先，包括 A 和 B 的整个题项序列将被加在一起以得到一个总分数。然后，计算 A 与总分数的相关和 B 与总分数的相关。假设这两个题项总分相关分别表示了 A 和 B 与题项背后潜在变量相对应的因子间的相关。如果只存在一个单一的潜在变量这一前提是正确的，那么在一个包含 A、B 和因子的路径图中，从后者到每一个前者之间都将存在路径（图 6.2 中的 a 和 b）。我们可以用题项总体相关来描述这些路径的值。根据这个路径图，A 和 B 之间的相关将是这两个路径的结果。计算 A 和 B 之间映射的相关只需要简单相乘。一旦被计算，这个映射的相关将与实际的 A 和 B 的相关进行比较。这个映射的相关可以和实际的相关相减产生一个残余相关。如果残余相关不等于零，那么这将表明，把一个单一的潜在变量当作 A 和 B 之间的共变的惟一原因是不充分的。

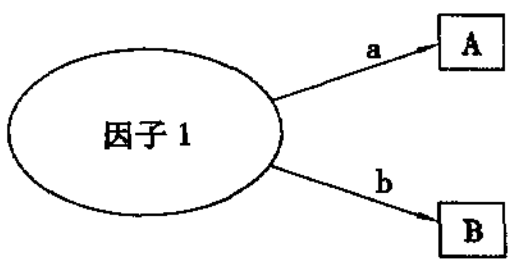


图 6.2　只包含两个题项的简化的单因子模型

可以同时对整个相关矩阵中的每对变量执行这种操作。不是只计算一个残余相关就结束了，而是计算整体残余相关矩阵（可称之为合理的残余矩阵），每一个残余相关表示了一个特定的题项对

之间的共变量，该共变量是一个单一的潜在变量所不能解释的。

继发因子(subsequent factors)

现在，用对待原始相关矩阵相同的方法从残余相关矩阵中抽出第二个与新的潜在变量一致的因子，这是可能的。再进行一次计算，能计算出题项和第二个潜在变量(例如，因子 2)之间的相关，并且根据这种相关也能产生一个相关矩阵。这些相关描述了在第二个因子被考虑后剩余题项之间的相关程度。如果第二个因子获得了抽取第一个因子后剩下的全部共变，那么这些映射值(projected values)将能与上面提到的残余矩阵的值进行比较。如果不是，那么就需要更多的因子来解释剩余的尚未归因于某个因子的共变。

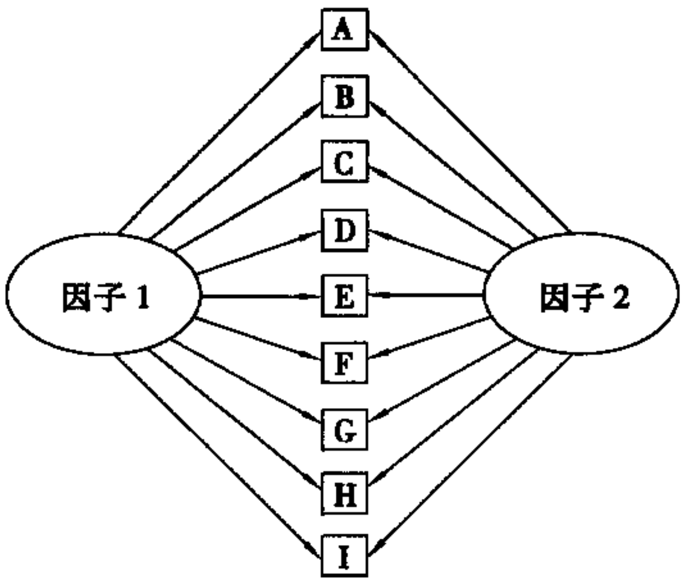


图 6.3　两因子模型

随着每个连续的因子被从先前交互作用而形成的残余矩阵中抽取出来，这个过程持续进行直至得到一个只包含小到可以被接受的残余相关的矩阵为止。此时，我们能确定，基本上全部重要共变已得到了解释并且不需要用更多的因子来解释。这个过程有可

能继续直到获得全部由零构成的残余矩阵为止。例如，在因素分析过程中，当抽取的因子的数量与题项的数量相等时，这种情况就会发生。换句话说，k 个因子的集合总能解释 k 个题项集合之中的全部共变。

决定抽取多少个因子

决定抽取多少个因子是一个棘手的问题（例如，Zwick& Velicer，1986）。进行因素分析的主要目的是，从一个大的变量集（项目）转移到一个能合理地抓住原始信息的较小集合（因子），即是说，精简信息。决定什么是"合理的工作"可以采用几种方法。

一些因素分析的方法，例如，那些基于最大似然估计（maximum likelihood estimate）和基于结构平衡的验证性因素分析程序（confirmatory factor analytic procedures，随后我们将讨论）的模型方法，采用的是统计标准。在本书中，术语"统计标准"判断某一个结论成立可能性是否小到足以排除其出现的随机性。这等于是执行一个测试来检验，在抽取每个连续的因子后，剩余的残余共变量在统计上是否远大于零。如果大于零，这个过程将持续至不大于零为止。依赖于统计的标准而不是一个主观的判断，是这些方法令人感兴趣的特征。当然，在量表编制中，这可能与其面临的目标并不一致，面临的目标确定了能解释题项间重要共变的一个小的因子集合。基于统计的方法的目的是寻找一个对潜在题项集的因子的详尽解释。如果存在不能由已抽取的因子解释的共变源，就必须继续抽取因子。量表编制者通常追求的是对因子进行谨慎的解释。即，在量表编制的过程中，我们通常想知道的是那些少量的、但又有重要影响的、支持题项集的共变源，因为我们不能找出所有的共变源。在编制一个量表时，通常会形成一个比期望要寻找的最终量表长得多的题项集。对那些已经确定对主要的因子没有贡献的题项，我们可以删除它们。我们的目标是：确定相对少量的、与潜在变量有强相关的题项。虽然熟练的数据分析者能通过使用统计标准的因素分析方法达到此目的，但对一个不熟练的研究者来说，采用其他更主观但不模糊的指导方法可能会做得更好一些。

这些相对主观的指导方法通常以一组因子能解释的原始题项的总变异的百分比为基础。这在本质上与以统计学为基础的方法是相同的。在采用非统计标准(如,不是基于概率)的情况下,数据分析者可以评估每个因子包含的大量信息,并在一个下降点达到回归点时作出判断。这是一个基于十分主观的标准而不是基于 p 值——一个统计学标准——对相关(例如,信度系数)进行解释的粗略类比。在足够的因子被抽取后,进行判断的两种广泛使用的非统计标准是特征值法则(Kaiser,1960)和碎石检验(scree test;Cattell,1966)。

一个特征值(eigenvalue)表示了一个因子所获得的信息量。对某些类型的因素分析方法(例如,主成分分析法,principal components analysis,在下一部分讨论)来说,题项集合的信息总量与题项数相等。因此,在一个有 25 个题项的分析中,将有 25 个单位的信息。每个单位的特征值与这些单位的某些部分一致。例如,在一个对 25 个题项进行分析的例子中,一个特征值为 5 的因子可以解释 20%的(5/25)总体信息,2.5 的特征值可以解释 10%,等等。如果一个题项集有 R 项,信息数与信息量的关系是 1.0 的特征值与题项间总变异的 1/R 相一致。也就是说,获得了 1.0 的特征值的因子(假定是主成分分析)获得了与典型的单个题项相同比例的总体信息。因此,如果因素分析的一个目标是获得少量的能充分获得在原始变量群中所包含的信息的变量,那么因子就必须比原始题项具有更大的信息负荷(loading)。因此,特征值法则(Kaiser,1960)认为,不必保留特征值小于 1.0 的因子(从而包含更少的信息量)。虽然排除这些因子的基本原理有道理,但对那些只稍微高于 1.0 的因子该怎么办?一个比典型题项多解释了 1%的信息的因子真的提供给了我们更多的信息吗?答案通常都是否定的,这表明特征值法则对于保留因子来说是一个太宽的标准。我相信这是传统的量表编制方法的一个较为普遍的问题。

碎石检验(Cattell,1966)也使用特征值,但是它是以相对值而不是绝对值作为标准的。它以联合连续因子的特征值而形成的地形图为基础。因为每个因子是在第一个因子抽取后,从先前因子抽取后的残余矩阵中抽取出来的(如前所述),因此每个连续因子

的信息量都少于前者。卡特尔主张,"正确"的因子数目可以通过寻找连续因子间信息量(即,特征值维度)的突然下降来决定。当绘图时,这种信息将形成以左边主要为垂直部分(表示大的特征值)过渡到右边相对水平的部分(表示小的特征值)为特征的地形图。他认为,图的右边的水平部分的因子是可以牺牲的。在外行看来,碎石是在山崩后在地面上收集到的小石子。这里采用这个术语,表明垂直部分是稳定的因子,而水平部分是碎石,或小石子,是必须丢弃的。在理想的情况下,在这个图形中,因子的信息量在某一点上会突然下降,存在一个陡峭的、从垂直到水平转移的一个清晰的"转折"(elbow)。

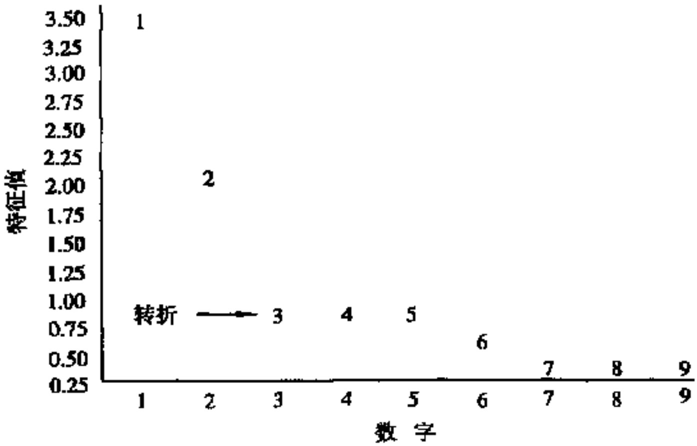


图 6.4 有明显转折的碎石图

卡特尔的标准要求保留存在于转折的上方的那些因子。有时,这种转折不是突然的而是缓慢的,是由在图的垂直区域和水平区域之间的几个因子构成的一条较为柔和的曲线。在这种情况下,应用卡特尔的碎石检验将需要慎重并且涉及甚至对主观标准的更大依赖,例如因子的可解释性。当与因子相关的题项彼此相似并且作为一个整体结构的指标具有理论和逻辑意义时,一个因子才被认为可以解释。

因子旋转(rotating factors)

抽取因子的目的只是确定用来测验的因子的适当数目。将信息处理成可以理解的方式不是抽取因子的目的。原始的、未经旋转的因子是一种无意义的数学上的抽象概念。作一个大致的类比,假设我被要求描述一下一个房间内所有人的高度。我决定随机选择一个人,例如,我选择乔(Joe)并测量乔的高度,然后以这个高度作为参照来描述其他人是高于这个身高还是低于这个身高。所以,一个人的身高可能是"乔的身高加 3 英寸"而另一个人的身高是"乔的身高减 2 英寸"。在这个例子中,所有的关于高度的信息都可以在我的数据表达中得到,但这没有被组织成最优化的信息组织方式。对人们来说,如果我把这些数据整理成更容易理解的方式,例如,把房间的每个人的高度用英尺和英寸来表示,那么人们就会很容易理解我的数据。因素分析与把已获得的数据以易于理解的方式进行转换相类似。

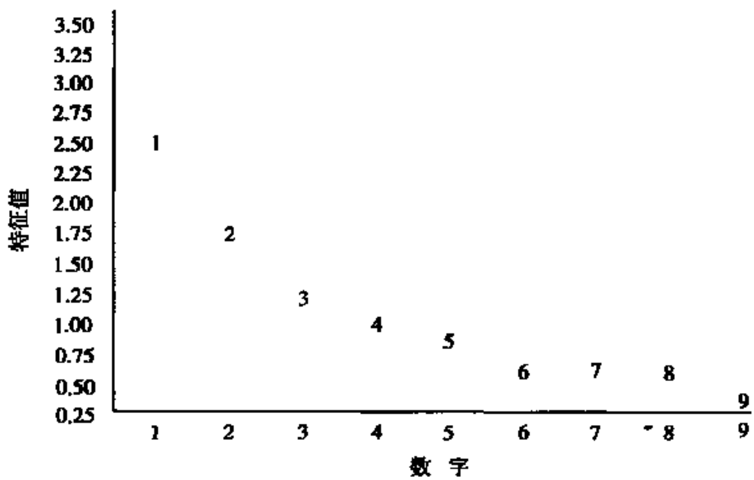


图 6.5 没明显转折的碎石图

在尝试解释因子之前——确定与因子对应的结构或潜在变量是什么,这依赖于与每一个因子相关的题项——通常需要执行因

子旋转。因子旋转通过确定一个由于单一潜在变量而使得特征突出的变量群来增加其可解释性，即它们全部只与一个或相同的因子之间有强的联系（并且在很大程度由此决定），这一点上题项是相似的。旋转和更大的解释性的产生并不是通过改变题项或题项间的关系来完成的，而是通过选择能更好地描述它们的方式来完成的。题项集中题项内部的相关模型与空间中的物理定位相似。如果两个题项越相关，那么代表这两个题项的标记就放得更近。如果我们对许多题项都这样做，那么，这些物理定位的标记将呈现一个表示变量间相关的模型（如果我们把自己限制在两个维度上，这一模型就很容易出现）。想象中的物理客体的定位是通过潜在的规则来决定的，而我们可以用另一种方式来思考题项间由潜在的因果变量决定的联系。

旋转类比 1

旋转怎样使我们看见变量间总是存在的但又不明显的模式？做一个类比，我们可以看一下一个布局非常好的墓地，例如阿林顿公墓，在这儿整齐地排列着统一的纪念碑。站在远处的某地往墓地看，你可能没有发现，这些墓碑和圆柱是沿一个有箭头的轴线方向排列的。如果所看见的线条并没遵循直线排列的任何一条自然轴线的话，这些墓碑似乎是随意放置的。当你改变观察点时，你可能会发现这些墓碑有特定的顺序。也许你向左或向右迈出一小步并使你的视线沿着一个对齐的标志观察这些墓碑，你就能发现它们的规则线。这时，每个标志都明显地和其他标志共享一条行（和列）。所有在一条线上的标志在某种程度上是相同的——它们拥有一个在先前的观察点不能发现的属性（即处在相同的线上这一关系）。因素旋转与为数据的组织结构提供一个观察点相类似，其目的是使题项间共享特征的方式变得明显。

值得注意的是，只要采用适当数目的正交参照线，不管多少条参照线都能定位客体。举一个二维参照线的例子，如阿林顿公墓（暂时忽略小山和山谷）就能阐明这一点。我能在公墓的任何一个定位点画一条线并确定与这条线正交的第二条线。用这两条线，我就能具体定位任意墓碑：我能说，“沿 A 轴 50 码，然后向右转 90°（事实上与轴 B 平行），并前进 10 码。”这将把你定位在一个具体的

点上。根据从墓地画出的任意两条正交线，使用其他合适的指导语，我能把你引入同一位置。所以，能适当地描述具体地点的定位线是任意的。任意两条正交线与其他的两条正交线在定位一个具体点时具有相同的信息功效。当然，这种假设成立的条件是：存在适当数量的线。在这个例子中，我简化这个公墓为一个二维空间。相应地，为了定位墓地内所有可能的点，两条线是充分而且必要的。如果我只有一条定位线，根据这一条定位线所确定的位置来引导你到达目的地，只能是一种偶然现象。因素旋转是一种用最直观的方式来确定适当数量线条的方法（在抽取因子的过程中决定）。

关于"最有效"的操作性定义指确定题项间固有的相似性（与墓碑共享一条共同的线相类似）并确定定位参照线，以便在分析的过程中使分析具有相同的维度（与沿着标记性的箭头的定位线类似，从这条线的第一个墓碑走到相同线条的最后一个墓碑）。为了达到此目的，我们也可以采用只有一个维度（例如，只沿标志线）的方式，即只用一条而不是两条线的方式来恰当地描述研究的步骤。虽然墓地有两个重要的维度，但是我们可以通过只采用其中一个突出的维度变量来描述墓地，例如沿着一行的位置。

类比通常是不精确的。在这个例子中，我们更多地强调如何根据一行的方向来描述墓地，而很少提到如何根据列来描述墓地。关键的问题是：定位适当的参照线可以使墓地的特征简化。下面的类比虽然也是不精确的，但我们能清晰地发现单一的维度是如何使特征简化的。

旋转类比 2

某些墓地（我期望这些例子不会让你感到太恐怖）是按下葬时间的早晚顺序排列的。一些欧洲古老的墓地是根据声望排列的，例如，最著名的和最虔诚的死者被安排到离教堂最近的地方。想象一下，一个墓地可能按上述所有的标准进行排列，表示埋葬时间的早晚的线条与教堂的墙平行并与表示声望的高低的线条正交。在描述之前，墓地的任意地点能通过参照任意两条正交线而被具体地定位。但是，请注意，如果我采用一条平行于教堂的线和另一条与教堂正交的线作为参照来描述墓碑，我能通过只使用其中的

一条线(沿埋葬时间早晚的前进线)或另一条线(沿墓地拥有者声望的高低的前进线)使大量关于墓地方位属性的信息得到简化。其中的一条线表示了坟墓之间的声望这一相同的维度。另外一条线表示拥有坟墓的时间长短这一相同维度(图 6.6)。

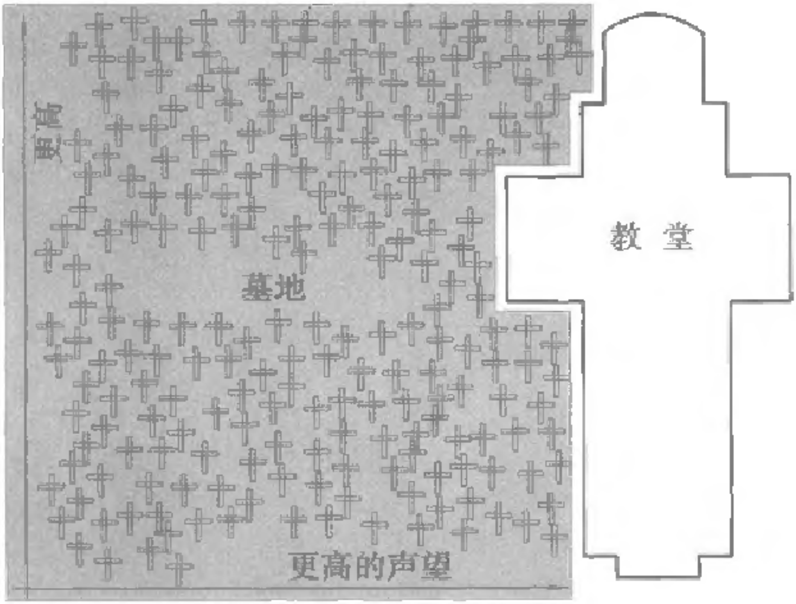


图 6.6 根据埋葬日期和死者的声望排列坟墓的假想墓地

如果一组旅行者想知道这个城市已产生的最有声望的公民的墓地可能在哪儿，我能告诉他们，直接沿着教堂的这条线走，死者的声望将逐步提高。相对而言，沿教堂的墙的那条线所代表的特征与声望无关。我也能指导旅行者沿平行于教堂的长轴的线条寻找这个墓地中最古老的墓。这时，离沿教堂的墙的这条线多远或多近与墓的古老程度无关。

相对于采用两条参照线来描述墓地的方式而言，只采用一条参照线也是一种有用的定位方式，因为它允许我通过只参照一个单一的值(如沿一个轴或另一个轴)而简化另一个也能决定墓地组织结构的变量的信息。如果旋转这两个轴，例如，顺时针旋转 43.5°，那么这将是一个不那么有效的参照线。如果某人刚开始的

时候采用这种参照线来定位某个墓碑的话，那么，当他想再次找到这个墓碑时，恐怕就不会很容易。虽然采用这两条轴线也能和早先的一对线条一样精确具体地定位某个墓碑，但它不是一个简便的有组织的定位方式。

这个墓地的类比与真实题项的因素分析的根本区别是，在我所描述的墓地中所有可能的地点都已被占用。也就是说，已有的坟墓表示了所有的声望高低和古老程度。在本质上，这个二维网格是被填满的。当然，在一个真实的墓地里，坟墓也不可能完全根据两条参照线的顺序组织起来。这可能是初步的工作并可能需要十分长的时间。当有更多的墓地时，就越需要在古老程度和声望之间进行协调。当对定位一个坟墓有更多的要求时，在墓地变得拥挤之前考虑这个问题可能要好一些。

人口密度不那么大的墓地加强了我们在量表编制中的类比，因为在后一种情况中，我们通常没有能表示问题中的两个维度的全部可能程度。当我们写下题项时，我们有意不使它们涉及研究中的多个变量。当一个题项明显能产生多个的潜在变量时，我们会丢弃它。结果，在两个变量的情况下，我们试图对我们希望测量的每个变量拥有强的、不模糊的题项。这将与一个高声望和十分古老，土质松软的坟墓定位的墓地版本相类似。

因此，如果我们想象我们的这个版本的墓地与试图表示两个潜在变量的题项集十分一致，那么它就只会有作为这两个维度中的每一个维度相对简单的例子的坟墓，即较早的埋葬和较高的声望。这种安排导致一组坟墓集中在教堂墙的一侧（例如，那些高声望的公民），而另一组明显地集中在教堂墙的正交线的末端，如图6.7所示。这时，坟墓同时用更古老和较高声望一起来定位（如图中显示的那样在墓地右上方的角落）。余下的坟墓中的任何一个都可以被明确地归类为属于这些墓群中的一个，而与其他墓群关系甚少。

在因素分析中，通过寻找导致每个题项主要负荷于（例如，与之相关）惟一一个因子，旋转就可得到明晰化。本质上，这是尝试着寻找一个与我们所描述的有选择性地埋葬的坟墓相类似的模式。这些题项通过强调所有题项都与单一因子相关，而以一种有

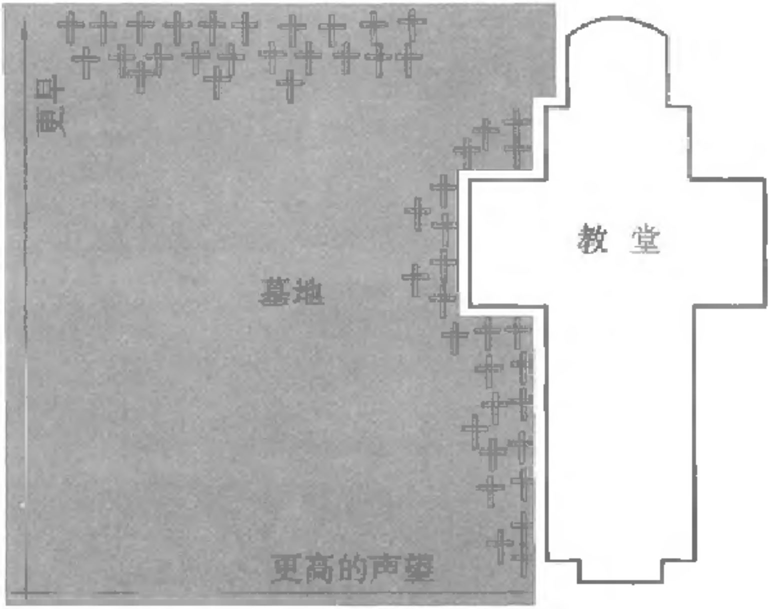


图 6.7　明显地分为非常古老和非常有声望两个墓葬群的假想墓地

意义的方式被赋予特征。每个题项在一个单一因子上有 1.0 的负荷并在其他因子上有 0.0 的负荷。这是一个完美的假想情况，被称为简单结构。在简单结构获得最有可能的近似计算法则时，旋转运算使用优化的数学标准。因此，如果假定某个潜在变量支撑某几个题项，那么一个成功的因素旋转的最终结果是：一个数据结构，该结构反应了那些共享有某些基本特征的题项的自然分类，可能是因为这些题项的一个共同潜在变量在起作用。同时，那些题项与定义其他题项群的任何特征之间很少有关系。

正交旋转与斜交旋转(orthogonal versus oblique rotation)

迄今为止的所有例证都以互相正交的基准线为基础。这与统计上要求的因子之间要互相独立一致，也就是因子之间是无关联的。因子之间被描述为互不相关的。当两条直线相垂直时，知道一条直线的知识并不能说明已经知道另一条线的信息。例如，知道一个人距离北方多远，并不表示能知道其距离西方有多远，因为两个方向是

互不相关的。同样地，在我们想象中秩序井然的墓地里，即使知道某一坟墓有多古老也不意味着知道已故者多有声望。

当两条基准线不互相垂直时，知道一条线的位置信息便能为另一条线提供一些位置信息。如果我们用一条想象中的大致穿越迈阿密到波士顿的线代替纬度，知道某个人处于该直线(或其平行线)的任一端，可为推断那人可能位于更北或更南端提供一些依据。沿着某一经度线编写的旅游指南和这条假定的线条是相关联的。

因子旋转也可以允许基准轴(和与它们相一致的因子)相互关联，而不必在空间上相互垂直。这种旋转被称为是斜交旋转而不是正交旋转。当潜在变量在某种程度上被认为相关时，斜交旋转可能是有用的。由于只有一个类别，简单结构的目的是使题项能被有意义地分类。也就是说，每个题项应该只和一件事相关，从而仅涉及一个因素。如果变量在某种程度上相关，但由于因素分析方法的限制，代表它们的因素被迫完全独立，那么便不可能实现该目标。也就是说，由于因子之间的相关，所以不止一个因子可能会和一些或所有的题项发生联系。我们将其近似地描述为简单结构时就会受到限制。

回到先前的合作者的品质这个例子上，如果责任心和可信性确实相关联，那么，和一个因子相关的题项可能也和另外的题项有相同的变异。然而，如果两种因子在某种程度上被允许相关联，那么其情形就大致和以下的推理类似：责任心和可信性被认为彼此相关。事实上也允许通过因子之间的相互关联来处理该论据。现在，撇开那不说，这些因子中哪个因子与问题中的题项具有最强的联系？因此，让因子本身之间互相相关使题项和其中一个或是其他的因子能被较为明确地分为一类成为可能，从而使我们更接近我们的简单结构的目标。即使给定的一对题项之间在因子水平上是相关的，题项也不必和两个因子都相关。

当因子被旋转至倾斜时，所失去的是不相关因子的精确性和简便性。不相关因子的一个非常好的特征是它们的组合效应是各自效应的简单相加。某个因子对某个具体题项进行解释的信息量可以被加到另一因子所解释的信息量里，从而获得两个因子共同

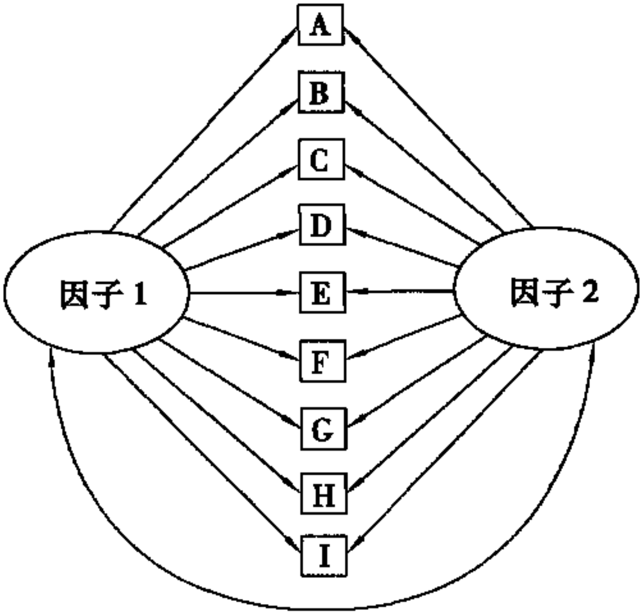


图 6.8　允许因子相关的两因子模型

解释的信息总量。有了倾斜因子,情况就不同了。因为它们相关,所以,这两个因子中就包含了多余的信息。对于一个与责任心和可信性都相关的题项,由两个因子共同说明的变异的数量小于各部分的总和。一个因子能解释的信息能和其他因子能解释的信息重叠。简单相加将使该部分重叠信息被加两遍,这不能准确地反映所有因于对这个题项的影响。

相关因子出现的另一个新问题是题项和因子之间的因果关系更具复杂性。当因子是互相独立的时候,一个因子和一个具体题项之间的联系是惟一的和直接的。因子水平上的变化将导致题项的直接的单一的因果关系的路径发生变化。但是,当因子之间是相关的时候,情况就不是这样的。例如,如果两个假设的因子都影响题项 A,并且因子之间相关,那么每个因子便对题项 A 产生直接和间接的影响。也就是,因子 1 影响因子 2,并通过因子 2 间接影响题项 A。这是除了因子 1 对题项的直接影响之外的又一影响。

当然，通过类似的过程，因子B通过和因子A的联系也直接和间接地影响了该题项。同样的直接加间接的影响也适合所有其他题项。因此，在谈及一个题项和一个因子之间的关系时，通常必须准确地限定包括或不包括那些间接的影响。此外，这种影响的模糊性能导致随后的混乱。

选择旋转类型

在实际情况中，对正交旋转和斜交旋转的选择应该存在一种或更多的考虑。其中之一是：一个人如何评价因子所代表的概念。如果理论足以支持相关的概念，那么它可能使得照着已有方法进行的因素分析(具体地说，是旋转)是合理的。因此，如果我们分析与责任心和可信性相关的题项，那么，使因子间相互关联便将最符合我们对这些概念的含义的理解。另外可能是，理论也许会认为因子之间是不相关的。例如，可信性和玩笑之间可能是相互独立的，因此可能会得出可信性与玩笑是不相关的结论。当理论不能提供强有力的证据，并且当量表还存在一些在此之前没有被研究过的表征和概念时，因子间相关性的大小可以作为指导。尤其是，斜交旋转可以被具体化，并且因子间的组合相关可以被检验出来。如果这些很小(如，小于0.15)，数据分析者就可以选择直角旋转。这是简单结构的一个近似的折中，但最终将导致更简单的模型。例如，一些题项可能表现出次要的负荷(即，在某个因子上有负荷而不是在某个因子上有非常强的负荷)，虽然这相对间接地、略微地增加了题项的负荷，但是仍然可以清楚地把每个题项和惟一的一个因子联系起来。因此，某个题项在被斜交旋转的3个因子上的负荷可能是0.78，0.16和0.05。当选择正交旋转时，负荷可能是0.77，0.19，0.11。虽然第二种范式比第一种稍稍背离了简单结构，但研究中的题项仍然能明确地和第一个因子联系起来。因此，这个例子中，选择更简单的(即，正交的)模型就没有什么损失。如果因子间高度相关，那么选择斜交的方案可以对近似的简单结构产生实际的改进。例如，和正交旋转一道获得的0.40的次要负荷可以和斜交的方案一起缩小至0.15。虽然这不是普遍的情形，然而只有对两种旋转方法之间的差异进行测量，才能够明确地说明它们在简单结构上的不同相似性程度。

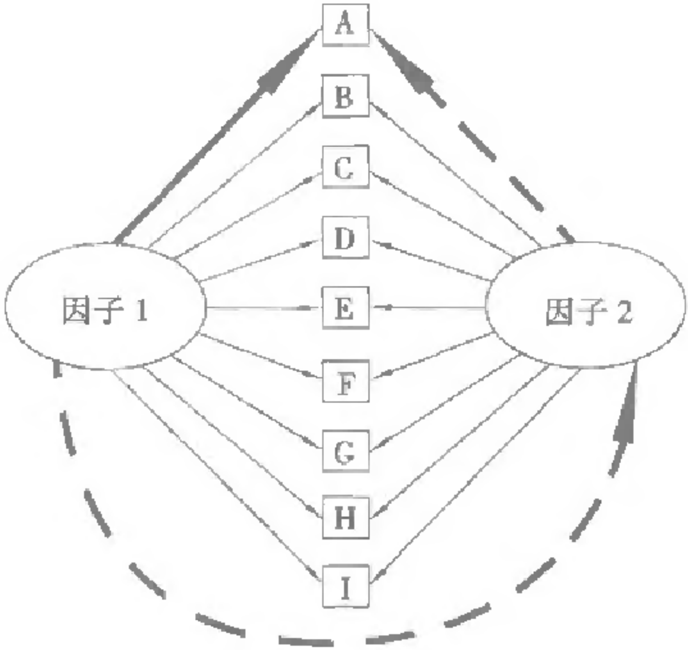


图 6.9 因为因子间相关，因子 1 同时直接影响（黑实体线）和间接影响（黑体虚线）题项 A

最后一个实际问题涉及两种因素之间相关性的大小以及在把两个因素结合为一个更大的相关之前的相关性到底要有多大。对于这个问题没有简单的答案，因为需要考虑题项和因子间的关系。但是，在某些情况下，即使两个因素之间具有高相关，斜交旋转的结果也可以表明一些题项对两种因素确实都有负荷。在那种情形下，抽出一个因子后，观察呈高相关的两个因子是否可以合并为一个因子就很有意义。例如，实际的数据将支持合并关于责任心和可信性的题项为一个因子，而不是分开它们。

因素的解释

在涉及责任心和可信性题项的例子中，我们曾假设我们准确地推断出潜变量是什么。在通常情况下，我们在随后的研究中将依靠因素分析所得出的与潜变量的特征有关的线索。这将通过检验那些最能说明每个因子的题项来完成，也就是在某个特定的因子上有最大的负荷量。有最大负荷的题项是那些最接近潜在变量的题项（同时也是最相关的）。所以，它们使我们能够讨论因子特征。当几个题项明确地把一个共同的变量和大的负荷（大于0.65）集中于相同的因子上的时候，检验是最容易完成的。回到同事的重要品质是什么这个例子中去，如果“聪明”、“固执”、“有教养”同时也许还有一个或两个其他的和智力水平相关的题项都在同一因子上有大的负荷，并且没有其他题项在那个因子上有大的负荷时，那么推断出“归因于智力的重要性”或某种等价的描述作为该因子的恰当表述将是相当容易的。

虽然在某些情况下为某个因子选择一个表述似乎明白易懂，但是确定一个名称和确定效度是不一样的。题项集是否继续作为所确定的名称的含义将最终决定效度。在解释方面，当因子解释相对少的变异，同时又有大量看似不同的题项负荷于因子上的时候，因素分析要尤其小心。如果分析得出了一个由看似不同的题项支持的因子，那么最好不要太过认真地把该因子视为潜在变量的指标。

在解释阶段值得提出的另外一点是：因素分析仅能发现用于说明被分析的题项之间联系的结构，并不揭示现象的本质。例如，某个试图确定人格的基本维度的研究者如果在量表中不包括关于外倾性的题项，因素分析便不可能获得外倾性因素。

有时候，包含一个特定的短语会产生一个关于概念上有意义的因子的错误表象。例如，当一些叙述用第一人称表达，而其他的不用第一人称时，可能会对这种观察到的联系模式作出解释。看看以下的假设题项：

1)我喜欢苹果。

2)桔子的味道很好。

3)水果中我更喜欢苹果。

4)很多人喜欢桔子。

5)我通常喜欢苹果。

6)桔子通常有令人愉悦的香味。

7)一个新鲜的桔子可以很好地款待他人。

如果奇数项负荷于一个因子,偶数项负荷于另一个因子,我们不会知道,基数项中的"我"这一称谓是否导致了两个因素,或者人们是否对所提到的两类水果表示了不同的态度。两种解释都似是而非且互相混淆。这是一种我们或许会或许不会把苹果比较成桔子的一种情形。

主成分与共同因子

有两种能巧妙地分析数据的方式:因素分析和主成分分析。一些作者认为,这些技巧从根本上说是相同的;而另外一些研究者则认为,它们是根本不同的。因素分析这一术语有时候同时包含这两种技巧,有些时候又被用来描述二者之间的对立。共同因子和主成分这些术语以一种较为清楚的方式表明它们分别源于因素分析和主成分分析。我们有基础来论述这两种方法的相同点和不同点。

主成分分析法(principal components analysis,PCA)可以用来分析一种或多种能从较大的题项集中获得大量信息的复合变量。此外,成分被界定为原始题项的加权和。也就是说,主成分是原始变量的线性转变。它们以实际数据为基础并源自实际题项。它们仅仅是实际题项中信息的重组。

共同因子分析(common factor analysis,CFA)也能用来分析一种或多种能从较大的题项集中获得大量信息的复合变量。但这些复合变量表示的是假设变量。因为它们是假定的,所以我们能获得对这些变量的全部的评估。共同因子是一种理想化的、假想

的结构(结构的本质通过检测其如何影响具体题项来推断),该结构可能使题项得以如实地回答。

成分和因子的异同

以上描述突出了成分和因子之间的一些不同。其中一点是:因素代表的是我们所评估的、理想化的、假设的变量,而成分及其结合在一起的信息是原始题项的可选择的形式。提取共同因子背后的实质是我们可以排除每个不与其他题项存在共同特点的题项。从因素分析的观点看,如果题项有信度,题项间没有共同的变异事实上是错误的。因此,在我们抽取共同因子时获得的混合信息是对理论上无误差(error-free)的潜在变量的估计。在这时,共同因子被理想化——它们是对决定一套题项集的无误差变量看起来是什么的评估。此外,因子"决定"了题项被如何回答,而成分由题项被如何回答来"定义"。因此,在主成分分析中,主成分是题项的最终产物,同时,在题项中获得的实际分数决定了主成分的性质。然而,在共同因子分析中,我们沿用的是一个理想化的、假定的变量,这个变量是得到题项分数的依据。一个因素是对假设变量的一种估计,并且表示的是题项分数的原因而不是结果。

两者的相似之处是什么?首先,二者间的计算差异很小。记住,共同因子分析的目的是估计某个理想化的无误差变量。但是我们必须从实际数据中产生这种估计。如我们所提到的,因素分析方法通常是以一个表示被抽取因子的题项之间的所有联系的相关矩阵为基础的。回顾一下第3章,我提出了在一个协方差或相关矩阵里,所有对角线外的值仅表示共同的或公共的变异。那么正如我所谈到的,相关矩阵只是方差—协方差矩阵的一种标准形式。它们之间的相关是标准的协方差并且是题项变异的标准联合。每个标准的题项方差表示了由某一题项证明的所有共同的或个别的变异性。为了产生一个理想化的、无误差的变量,必须除去沿着相关矩阵中主对角线所包含的题项方差的特殊变异部分。更具体地说,每种联合必须用共同的估计(communality estimate)——包含在因素分析中的某一特定变量和其他变量的共同变异的小于1.0的一个近似值——来替换。例如,如果我们估计

某一特殊变量在相关矩阵中占其他题项的总变异的 45%，那么我们将能确定共同的估计是 0.45，并用它代替用 1.0 表示的题项总变异。我们将对每个变量做这样的处理，即用共同的估计代替每种联合(通常，共同的估计可以通过问题中的变量在剩余变量中的回归来获得，即把这种回归获得的复合相关的平方，R^2，作为共同的估计)。这个过程可以产生一种可变的相关矩阵，该矩阵被用来提取公因子而不是主成分，如表 6.1 所示。

表 6.1 主成分分析和共同因子分析的相关矩阵

1.0	0.70	0.83	0.48	0.65
0.70	1.0	0.65	0.33	0.18
0.83	0.65	1.0	0.26	0.23
0.48	0.33	0.26	1.0	0.30
0.65	0.18	0.23	0.30	1.0

0.45	0.70	0.83	0.48	0.65
0.70	0.52	0.65	0.33	0.18
0.83	0.65	0.62	0.26	0.23
0.48	0.33	0.26	0.48	0.30
0.65	0.18	0.23	0.30	0.58

注：左边的相关矩阵采用的是主成分分析，保留的单元在主对角线上。右边的相关矩阵采用的是共同因子分析。在主对角线上的是共同的估计而不是单元。

用共同的估计取代联合是区分共同因子提取和主成分提取的惟一计算差别。

“原因与结果”这个问题是怎么一回事呢？这是不是我们在分析观察到的题项分数时同时获得因子和成分的一种情形呢？正是。正如共同的估计所表明的那样，题项之间在经验上的关系最终形成公因子的基础。当然，成分也同样如此。因此，在计算上，两者都以实验数据为基础。而且，大多数分析者把使成分和共同因子概念化作为理解题项集潜在变量的方法。也就是说，成分和因子这二者通常被认为揭示了在题项集上所观察到的分数的原因。事实上，成分分析和因素分析通常可交换使用。大多数情况下，题项共同拥有一些有意义的东西，不同的方法得到的是同样的结论。因此，虽然这两者间有技术上的相似点和不同点，但是二者间的不同常常被忽略了。

不过，主要不同的一点是：成分和因子在解释变异的性质上是不同的。前者说明了原始变量之中总变异的某一特定部分，而后

者则说明了原始变量间共有的或公共的变异。如果减小相关矩阵的对角值，那么，像提取共同因子时所做的那样，变异的比例表达式的分子分母也会随之减小。但是分母减小的程度更大，因为这涉及相关变量的特殊计算。结果，由一系列对比性成分和因子所"解释的变量比例"是不等价的或在概念上不等同。因子解释有限的方差中(例如，共同方差)较大的比例，而成分解释总方差的较小的比例。当讨论因素分析的结果，报告因子所解释的方差比例时，弄清楚分析的类型(主成分或共同因子)，从而弄清楚被解释的方差类别(公共的或全部的)是很关键的。

两种分析类型间值得注意的另一不同是：在一些统计包里，提取主因子而不是成分而得到的结果将明显是毫无意义的。在两种分析中，被解释的方差的累积量会随着每个连续因子或成分的提取而增长。有共同因子存在，这个比例通常超过 1.0，并且，当考虑到连续因子时，这个比例会持续增加，之后，似有魔法般地，当第 k 个(即，最后可能)因子被提取时，又正好回到 1.0。这虽然看起来奇怪，但它仅是一种手工计算方法，是可以被忽略的。如果数据分析者用理性的标准来决定提取多少因子，那么，所选择的数目通常会超过在抽取序列中这种异常现象出现的那一点。但是，用已选择的因子数目来有效地解释原始题项中的所有公共方差(例如，100%)是可能的。

验证性因素分析

另一个因素分析方法的差别在于是探索性(exploratory)还是验证性(confirmatory)。这些术语原本是指数据分析的目的而不是计算方法。因此，同样的分析可能用相同的题项集去确定它们潜在的结构是什么(探索)或者是基于理论或先前的分析结果去确定一个预先假定的关系模型是否正确(验证)。随着使用频率的增加，这些术语现在被用于区分不同种类的分析工具而不是用于区分不同的研究对象。当人们在使用验证性因素分析这一术语时，他们通常谈论的是基于平衡结构模型(structural equation model-

ing,SEM)的方法。即使这些方法应该用于验证性的而不是探索性的情况,但标准的因素分析技术能够用于这两种情况。因此,验证性并不必然是以 SEM 为基础的。

然而,比起传统的因素分析方法,以 SEM 为基础的方法在某些情况下能够表现出实实在在的好处。之所以表现出这些好处,是因为 SEM 是一种非常灵活的结构。传统的因素分析方法要求的条件,比如题项之间的误差相互独立,在 SEM 的使用中,可以有选择性地进行改变。当然,传统的因素分析方法对数据分析者的大部分限制是要求因子间相关或者相互完全独立。但如果理论表明有这样的一种模型存在的话,以 SEM 为基础的方法能够把相关和不相关的因素混合起来研究。

就像先前所述的那样,以 SEM 为基础的方法也能够为评定实际数据在多大程度上符合特定的模型提供一种统计的标准。恰当地使用它,它便是一种很有用的工具。然而,有时它会导致过多的因素分析。提取更多的因子经常会提高一个模型的适用性。提供一个严格的统计意义上的标准会模糊这样一个事实:一些统计上显著的因子只解释了非常小比例的变异。尤其是在量表编制的早期,这可能与研究者的目的相反,研究者关心的是找到极少量包含大多数信息的变异的因子,而不是能解释大量可能变异的因子。

以 SEM 为基础的方法是用共同的方式测试多个模型并比较它们对数据的适合程度,这是一把双刃剑。再次重申,如果谨慎地使用 SEM,这能够成为有价值的工具。反之,如果使用不慎,这就会出现几乎没有什么理论意义但却更具统计意义的模型。例如,取消关于误差相互之间没有相关这一限制,也许模型产生不了多大的价值,但是这个模型也许会在统计方面超过有限制性的模型。一位研究者也许决定忽视这些细小的相关以利于更简单的模型,而另一个研究者则会因为统计标准而拒绝更为简洁的选择。另一个例子是:一个把两个相互区别的但是存在高相关的因子(也许就像责任和信任)分开的模型可能比把这两者联系起来的模型更具适应性。如果这两者的相关非常高而把这两者分开就武断了。例如,假设同一结构的两个指标之间的相关为 0.85,通常这被认为是这两者等值的好证据。但是,把相关为 0.85 的两个因素分开的模

型比把这两者合为单一的因素的模型能更好地拟合数据。

这些评论并不是打算说以 SEM 为基础的验证性因素分析方法不好。这些方法的出现为理解不同的测量问题做出了大量贡献。然而，我发现这些方法内在的灵活性有很大的做出错误决定的可能，尤其是数据分析者对这些方法不熟悉的时候。除了主成分分析方法以外(这里的因子是题项的线性联结)，没有因素分析方法能产生惟一的正确解决方法。这些方法只能产生似是而非的解决办法，这类方法有很多。不能保证在统计意义上胜过简单方法的那些复杂方法在反映真实时更为精确。它可能是更精确也可能不是。所有的因素分析方法的共识是需要作出最佳决定。分析只是指导决策过程并为决策提供证据。在我看来，它们不应该代替调查者做出决定。同样，准确地在正式的因素分析的书面报告中描述决策、统计或其他的检查是非常重要的。

最后需要注意的一点是，某些领域的研究者(例如，人格研究)认为，在阐明模型的良好适应性方面，传统的因素方法比采用统计标准的方法具有更强大的解释性。例如，索塞尔和哥德伯格(Saucier & Goldberg, 1996)认为："因为解释性的因素分析提供了比验证性因素分析更严格的重复，前者比后者更为常用。" (p. 35)一个推论是：如果在不同情境中提取不同的个体样本数据，那么使用解释性方法必然产生同样的因素分析结果。虽然这些结果重复出现的可能性非常小。记住，在这种情况下，使用以 SEM 为基础的方法，数据分析者需要预测变量和能够与实验数据拟合的计算机程序得出的结果之间的关系。换句话说，实验数据是否与模型拟合强有力地预测了计算机程序将要得到的结果。相反地，在不能作出预测的情况下，重新考虑先前的因素结构并反复进行解释性的分析，能获得对实验结果具有强的说服力的解释。

量表编制中因素分析的使用

下面的例子将使在本章中所讨论的一些概念更加具体。我与我的同事(DeVellis, Devellis, Blanchard, Klotz, Luchok, & Voyce,

1993)编制了一个问卷来评估父母对于影响他们孩子健康的人和事的看法。虽然这一量表总共有 30 项并且评估了这些看法的几个方面，但是我只讨论其中的 12 项。

A. 我能够影响我孩子的幸福感。

B. 我的孩子是否能避免伤害只是运气问题。

C. 在决定我孩子的健康状况方面，运气起着重要作用。

D. 我能在防止我的孩子受到伤害方面起很大作用。

E. 我能在防止我的孩子生病方面起很大作用。

F. 我的孩子是否能避免生病只是运气问题。

G. 我与孩子在家里做的事会成为我的孩子幸福感的重要方面。

H. 我的孩子的安全依赖于我。

I. 我能做很多事帮助我的孩子活得更好。

J. 我的孩子的健康是一大笔财富。

K. 我能做很多事帮助我的孩子强壮、健康。

L. 我的孩子是否健康或生病只是运气。

我们一共对 396 位父母进行了调查并对结果进行了因素分析。因素分析的第一个步骤是决定这些题项中包含了多少个因子。SAS 用来进行因素分析，点状图是必要的。SAS 打印出来的点状图形式，如图 6.10 所示。注意，12 个因子(与题项数相同)都被标明了。而且，因素分析中有两个因子被定位于点状图上边位置，其余的沿着点状图的底端分布。这就有力地证明了这两个因子可以说明这些项之间的许多变异。

在决定有多少因子需要保留后，我们重新指定了两个因子并进行了方差最大(varimax)旋转(直角的)。如果我们无法接近简单结构的话，我们本来或许可以进行斜交旋转来提高题项与因子之间的适合度。然而，在这个情况下，直角旋转产生了具有很大意义的题项集和强有力的、明确的负荷。

这可以从下面的因素负荷表中明显地看出，在这个表中每一

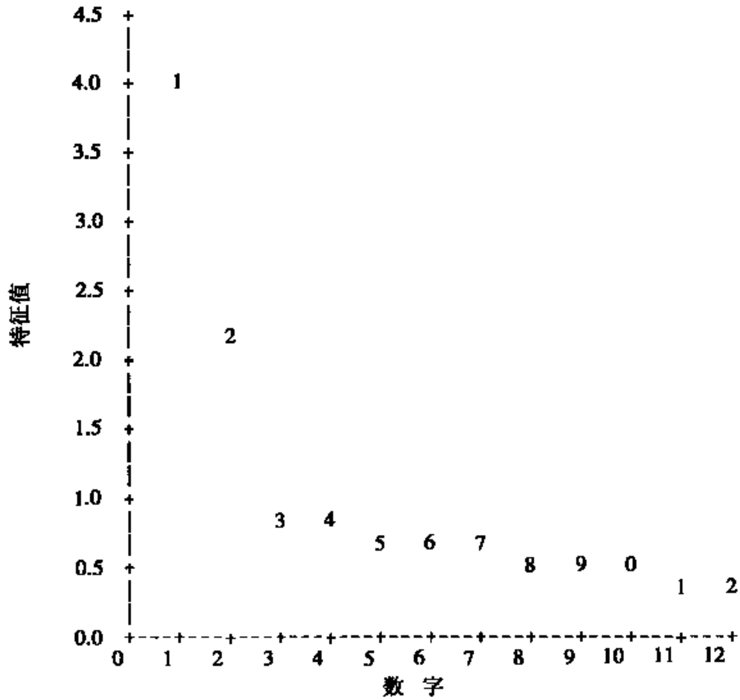


图 6.10 已选题项的因素分析的碎石图

排中都包含了一个给定的题项在两个因子方面的负荷。在 SAS 中可以进行的另一个选择是重新排列表中的题项，即那些在每一个因子上有着高负荷的题项可以被分在一组。

在表 6.2 中，因素负荷大于 0.5 的被写成黑体。每个因子由负荷最重的题项来定义(例如，那些黑体)。通过参照这些题项的背景，能辨别每个因子表示的潜在变量的性质。在这个例子中，所有在因子 1 上有强的负荷的题项关心的是父母是否在儿童获得安全和健康方面具有影响。另一方面，主要负荷集中在因子 2 上的题项，关心的是运气或命运对儿童健康的影响。

表 6.2　两因子上的题项负荷

	已旋转的因子模式	
	因子 1	因子 2
项目 I	**0.78612**	−0.22093
项目 K	**0.74807**	−0.18546
项目 D	**0.71880**	−0.02282
项目 E	**0.65897**	−0.15802
项目 G	**0.65814**	0.01909
项目 A	**0.59749**	−0.14053
项目 H	**0.51857**	−0.07419
项目 F	−0.09218	**0.82181**
项目 J	−0.10873	**0.78587**
项目 C	−0.07773	**0.75370**
项目 L	−0.17298	**0.73783**
项目 B	−0.11609	**0.63583**

这两个同质的题项组能被进一步检验。例如，能计算每组的 α 系数。利用 SAS 计算的每组 α 系数的结果在表 6.3 中。

两个量表具有可接受的 α 信度系数。注意，SAS 程序能计算非标准和标准题项的 α 系数。后一种计算等于是使用以相关为基础的 α 系数公式。对于这两个量表，两种计算 α 系数的方法得出了十分相似的值。注意，对量表来说，减少题项也不能增加 α 系数。减少一个题项，例如，在量表 1 中减少 H 题项和在量表 2 中减少 B 题项，减少题项后的 α 系数几乎和整个量表获得的 α 系数一样高。当然，保留这些题项能提供一点额外保障，因为它能使一个新的量表的信度不会下降到低于可接受的水平，并且不会在实质上增加量表的长度。

表 6.3　对两个不同的题项系列，所有题项和 $k-1$ 题项联合体的 α 系数

克龙巴赫 α 系数				
对原始变量：0.796472　对标准化的变量：0.802006				
删去的变量	原始变量		标准化变量	
	与总体的相关度	α	与总体的相关度	α
项目 I	0.675583	0.741489	0.676138	0.749666
项目 K	0.646645	0.748916	0.644648	0.755695
项目 E	0.545751	0.770329	0.535924	0.775939
项目 D	0.562833	0.763252	0.572530	0.769222
项目 G	0.466433	0.782509	0.474390	0.787007
项目 H	0.409650	0.793925	0.404512	0.799245
项目 A	0.437088	0.785718	0.440404	0.793003

对原始变量：0.811162　对标准化的变量：0.811781

删去的变量	原始变量		标准化变量	
	与总体的相关度	α	与总体的相关度	α
项目 F	0.684085	0.748385	0.682663	0.749534
项目 C	0.596210	0.775578	0.594180	0.776819
项目 J	0.636829	0.762590	0.639360	0.763036
项目 L	0.593667	0.776669	0.592234	0.777405
项目 B	0.491460	0.806544	0.493448	0.806449

通常，在编制量表时，对于可以进行因素分析的量表来说，可以采用一些说明来对量表的应用过程中可能发生的错误进行预防。例如，利用一个独立样本来重测量表的信度是十分重要的。事实上，在一个独立的样本上重测整个因素分析过程，以阐明所获得的结果不是一个偶然出现的结果，这可能是有用的。

样本大小

在原始分析中，样本大小至少部分地决定了因素结构重测的可能性。在通常情况下，在大样本因素分析中出现的因子模式将比在小样本中出现的因子模式稳定。不可避免地，问题出现了，“多大才算是足够大?”这很难回答(MacCallum, Widaman, zhang, &hong,1999)。和许多其他的统计程序一样，被试的相对数量和绝对数量都需要考虑，但诸如题项共同因子等因子仍起作用(MacCallum et al. ,1999)。需要提取因子的题项数目越大并且期望的因子越多，分析中包含的被试数量就越大。基于这样的事实，问题的关键是，寻找一个被试与题项的标准比例。当然，如果得到的样本足够大，被试与样本的比例将减少。对于一个有 20 个题项的因素分析来说，100 个被试可能太少，但对于 90 个题项的因素分析来说，400 个被试可能足够了。廷斯雷(Tinsley,1987)主张大约每个题项有 5～10 个被试这个比例，最多大约 300 个被试。他们主张，当样本大到 300，这个比例将减少。在同一论文中，他引用了其他规则，根据卡蒙瑞(Comrey,1973)，100 个被试太少，200 个被试比较好，300 个被试恰好，500 个被试十分好，1 000 个被试极好。卡蒙瑞(1988)声称 200 个被试的样本大小对于不超过 40 个题项的大多数一般因素分析来说是足够的。虽然样本大小和因素分析的有效性之间的关系比这些简略表示的规则复杂得多，但在大多数情况下仍能很好地为研究者服务。

在量表编制中采用较为适度的样本(例如，150 个被试)来进行因素分析可能不常见。当然，在因素分析中采用较大的样本来增加结论的推广性，这一观点是很好接受的。当然，在一个分开的样本上重测因素分析的解决方案可能是阐明其推广性的最好方法。

结 论

因素分析是量表编制中的一个基本工具。它允许数据分析者决定支持题项集的潜在变量的数目并正确地执行计算克龙巴赫(Cronbach)α系数的程序。另外,它也能让我们洞悉支撑我们题项的潜在变量的实质。

7

项目反应理论概述

An Overview of Item Response Theory

项目难度(item difficulty)

项目区分度(item discrimination)

假阳性(false positive)

项目特征曲线

IRT 的复杂性

何时使用 IRT

结　论

项目反应理论(IRT,item response theory)是经典测量理论(CMT,classical measurement theory)的一种替代方法,CMT 也叫经典测验理论(CTT,classical test theory)。IRT 近年来已经受到越来越多的关注(参看 Hambleton, Swaminathan, & Rogers, 1991; Embretson & Reise, 2000)。经典测量理论的基本思想是观察分数是被试的真实分数加上误差的结果。误差并不作进一步的区分,比如通过时间、背景或题项来进一步区分。相反,所有误差都用误差这一单独的术语来表示。IRT 方法则更好地区分了误差,尤其是需要考虑题项特征的时候。

尽管 IRT 主要应用于能力测验(诸如学业倾向测验),但在其他的领域,IRT 也得到了广泛应用。经典测量理论主要关心组合,更具体地说,是量表的组合。但 IRT 主要集中于每个题项和它们的特征。在 CMT 中,题项在某种意义上是得到结果的途径。也就是说,它们是同一潜在现象大致相等的指标,它们通过聚合为一个量表来获取力量。一个量表的信度靠冗余的题项来增加。在 IRT 中,每个题项与研究的变量(通常也称作属性)的关系都会得到评估。信度不是靠冗余题项增加,而是靠确定更好的题项来提高。更多基于 IRT 的题项通常会根据可被区分的属性联合体而增加节点数目,但并不能按照我们先前设想的方式增加信度。例如,给一个数学测验增加更多的难题只能向上扩大它的使用范围,而不一定对它的内部一致性有任何影响。在 CMT 中,编制量表的要求是题项具有共同的因果关系,因而彼此相关。而且,一个量表只能有单一的潜在维度。IRT 也具有这一特征,即将被组合在一起的题项必须共享单一的潜变量。因此,CMT 题项是冗长的,实际上这些冗余题项是量表信度的重要部分。然而,尽管 CMT 题项被编制的彼此很相似,并且这些题项以同样的方式来反映潜变量,编制的 IRT 题项却反映了属性的不同程度或水平。精度不仅靠聚合起来的冗余题项来提高,更要靠具有特殊的可论证特征的非冗余题项来提高。我们将在本章的稍后部分讨论这些特征。

因为项目反应理论源自能力测验,能力测验包含的题项通常和该领域的内容相关。也因为能力测验的题项通常以正确或不正确来分级(即使它们的最初形式涉及两个以上的反应选项),IRT

的经典应用和实例涉及的题项都呈现出两种状态中的一种(例如,“通过”或“失败”)。尽管没有理由说明为什么源自该理论的方法不应扩展到(事实上他们已经做了)具有其他反应形式(例如利克尔特量表)并适合其他内容领域的题项,但是通过对这种类型题项的讨论发现,IRT 是最简单的。

IRT 的目标是使研究者能够证实题项的某些特征,这些题项与完成它们的人无关。这与物理测验相似,物理测验能够评估一个物体的一种属性(如长度或重量)而不考虑它的特殊性质。例如,无论称什么,20 磅都表示同样的意义。这样,一个普通的称就能测出关于一个物体某一特殊属性的信息(如重量),而不管被称量物体的本质如何。IRT 期望用问卷题项达到同样的目的。

IRT 事实上更像一种模型而不是把一套单独的程序特殊化的理论。区分不同 IRT 模型的一种重要方法是看它们考虑的题项参数的数目。近年来一种常见的模型是三参数模型。毫不奇怪,该模型专注于题项表现的三个方面。这就是题项的“难度”、“区分度”和“灵敏度”。IRT 家族中一个很早但仍然流行的模型是拉希模型(Rasch Modeling:Rasch, 1960; Wright, 1999)。该模型只测量难度参数。

项目难度(item difficulty)

尽管这个术语明显是从能力测验沿袭下来的,但它所代表的概念却有更广泛的应用。项目难度指的是被测量的属性的水平,该属性与从“失败”到“通过”该题项的转换相联系。我们大多数人都看过描绘狂欢节的老电影或表演某种力量技艺的娱乐园。测量装置包括一个使重物沿其滑动的竖直滑轨,在滑轨的顶端是一个响铃。最初,重物在滑轨的底部,并放在一种用作跷跷板的木板的一端。“被试”用一个大木锤敲击跷跷板与重物相对的一端,这样就使重物沿滑轨向上弹起。他们的目的是用足够的力来推动重物,使其撞击响铃并敲响它。相对于我们的目的,我们可以把整个装置想象成“题项”。

项目难度是“被试”为了“通过”题项（如敲响响铃）必须拥有的力量总和（更准确地说是他或她必须传递的力）。显然，可以构造不同难度水平的题项（如，更难的题项具有更长的滑轨或更重的重物）。然而，确定一个特殊装置的校准难度应该是可能的，该装置独立于偶然挥动木锤的人的任何特征。

因为这个“题项”是一个物理实体，所以以合理的精度确定要使铃响需要多少力是相当容易的事（忽略敲击时相对于敲击位置的细微差别）。所以狂欢节的组织者可以放置一个 10 磅或 100 磅的装置在玩游戏的人中得到一个高或低的通过率。每个装置可能适合于不同的人群，如儿童参加学校展览会，成人参加健身训练营。

我们可以用相似的方法来表现问卷题项的特征。例如，设想一个测量抑郁的题项，可以把题项编制的相对“简单”或相对“容易”。首先，只需要通过适当数量的具有抑郁特征的题项（一星期至少经历一次特殊感情可能被定义为评价回答者的指标）。例如，诸如“我对我不得不做的一切感到沮丧”这样的题项很可能在这个意义上是“简单的”。但是个人一周一次或多次具有这种感觉的可能性不会取决于被提问者是谁。例如，如果我们向临床抑郁患者提出该问题，我们很可能会发现他们比普通人群具有更大的样本比例“通过”该题项。确定项目难度的目标就是在绝对意义上建立通过题项所要求的特征的多少。如果能够做到这些，那么一个人通过题项就具有了关于抑郁水平的稳定的意义，而与这个人是谁或者所研究样本的平均抑郁水平无关。换句话说，在描绘一个人的特征时不仅仅参照一个特殊样本的标准而且还参照独立于任何特殊样本的衡量标准。

项目区分度(item discrimination)

IRT 关心的第二个参数是题项把一个反应按“通过”或“失败”明确分类的程度。换句话说，对一个人是否真正通过或失败区分的越明白，问卷题项的区分度就越高。用我们的狂欢节响铃类比，

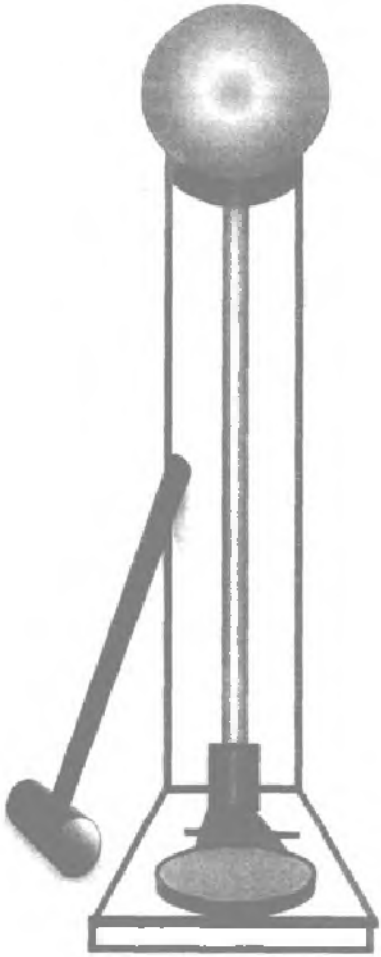


图 7.1　一个假定的通过使用足够的力用锤子敲打衬垫使铃响的装置来测量强度

有可能在偶然的情况下，重物没有接触到响铃，但铃却响了。观察者所认为的响铃是否真的响了也可能是不一致的。有人可能听到一个微小的响声而有人可能什么也没听见。当用力推动重物使其触及响铃但没有使所有人都同意这是一个清晰的铃声时，该装置则提供了一个模棱两可的信息。从另一个角度看待这种模糊性

(Ambiguity),用同样的力敲打多次也可能使观察者在某一时刻认为铃响了而另一时刻又认为铃没响。一个稍微大一点的力将持续产生一个清晰的响铃声,而稍微小一点的力却可能使人相信铃没有响。所以该装置的模棱两可的范围应该被减少。一个替代装置可能具有不同操作方式并且产生更少的模糊结果。例如,在重物撞击响铃的同时闭合一个延迟电路,使一盏灯发光并保持光亮直到重新开始测试。如果设置的好,这样的装置很可能在一个模糊相当小的范围上得到一个一致的结果,这样将会比标准装置有更好的区分度。相反,如果一个装置根本没有响铃,要求观察者看见重物超过紧挨在滑轨上的预先标记的一根线就举手来代替,这样的装置可能得到更模糊的结果,并且区分度也更差。所以,一个具有良好区分度的装置或题项,对于所研究的现象来讲,可能产生模棱两可的结果的范围只占很小一部分,一个低区分度的装置或题项具有更大的模糊区间。

假阳性(false positive)

IRT 的第三个参数是假阳性。假阳性是指一个反应显示某些特征或属性的水平存在而实际上它不存在。这里,我们需要再做一次狂欢节类比。你可能曾经见过这样的小屋,在一个水箱上面,一个人坐在一个塑胶玻璃保护屏后的一个平台上,平台连接在一个杠杆上,杠杆的一端有一个靶子。

比赛者向靶子投掷棒球,如果击中,就会使平台坍塌,并且使坐在平台上的人落人他或她身下的水箱中。我们可以把这个装置想象为测量投掷准确性的一个"题项",把平台上的人落入水箱中作为题项的"通过"(现在你应该能够描述装置的变化怎样增加或降低装置的难度和区分度)。通过这种特殊的装置,我们可以想象"假阳性"是怎样出现的,也就是说,一个实际上没有能力的被试是怎样使坐在水箱上面的人浸没而获得一个"通过"分数的。一种方法可能是被试胡乱的扔球但球碰巧击中靶子(毕竟,它必须要击中某处)。或者另一种可能,该装置可能发生故障,平台自动坍塌。

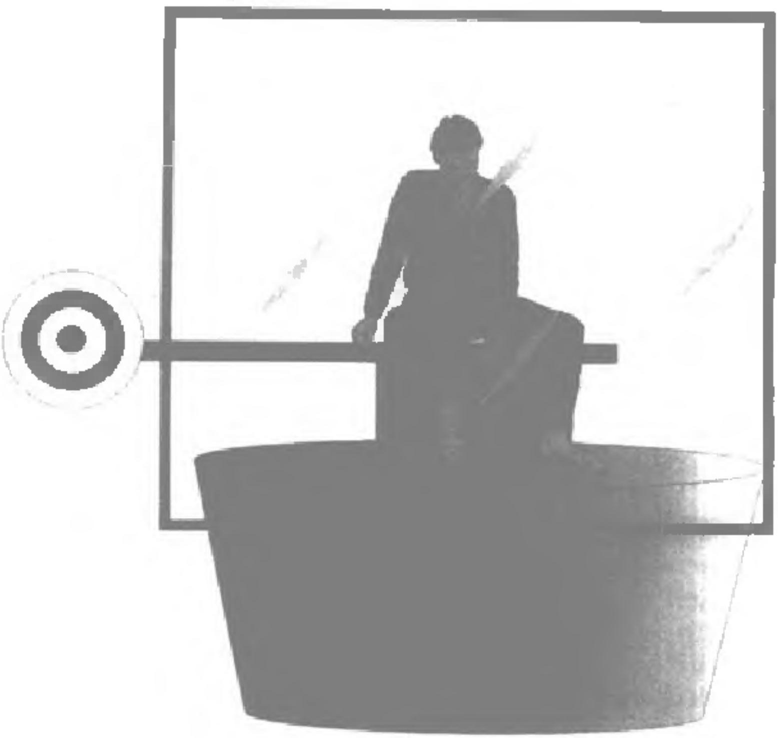


图 7.2　一个通过撞击目标球使平台下降并使坐在上面的人陷入有水的容器来精确测量投掷运动的假想装置

在这些情形下，玩家将会因为某些不相关的环境因素而不是能力因素而通过该题项。这样，即使一个人在投掷准确性上很差或没有能力，也可能通过该测验。在能力测验中尽管不是真正知道答案，但能成功的猜测正确反应，这种情形非常普遍，它的结果就是“假阳性”（在猜测或其他类型的假阳性几率很小的测验中，例如用天平测重量，通常二参数模型就足够了）。

这三个题项参数中的每一个——难度、区分度和假阳性——与测量误差都有相当明显的关系。如果①一个题项的难度不合适，②在通过和失败之间的模糊区间太大，或者③一个特征不存在而题项却表明它存在，那么该题项就有错误的倾向。IRT 为量化

题项的这三方面的性能并且为在已知背景下选择性能好的题项提供了一种方法。

项目特征曲线

这种量化结果随后被概括为一种项目特征曲线(ICC,item characteristic curve)的形式,它以图表的形式反映出题项的特征。通常 ICC 大致呈 S 型,并且曲线的不同部分揭示了关于所研究的 3 个参数中每一个参数的信息。

图 7.3 展示了 ICC 的外形特征。X 轴代表所测量的特征或属性的强度(例如,知识、力量、准确性、压力、社会期望或者可能任何其他可测量的现象)。Y 轴代表问题的通过率,它是以观察分数中失败和通过的比例为基础而得到的。如果我们比较代表两个题项的曲线图,理解怎样用 ICC 来评估题项的质量实际上非常简单。

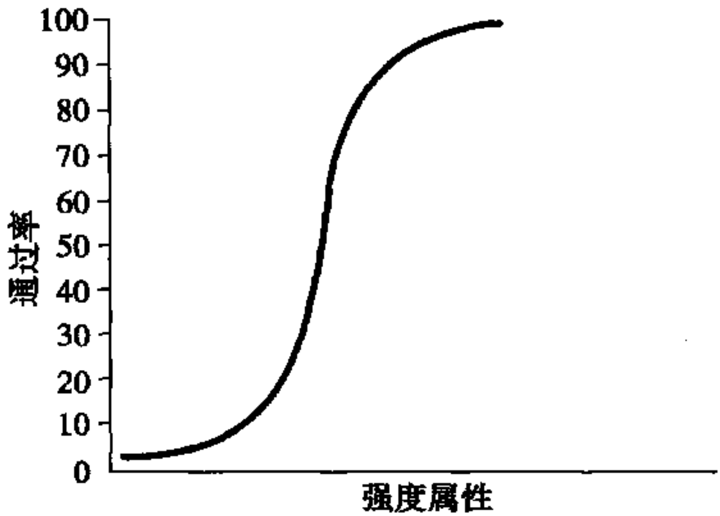


图 7.3 项目特征曲线(ICC)的一个例子

图 7.4 通过两条曲线阐明项目难度。注意,各个题项达到 50%的通过概率的点是不同的。对浅色的曲线,该点更靠右边。

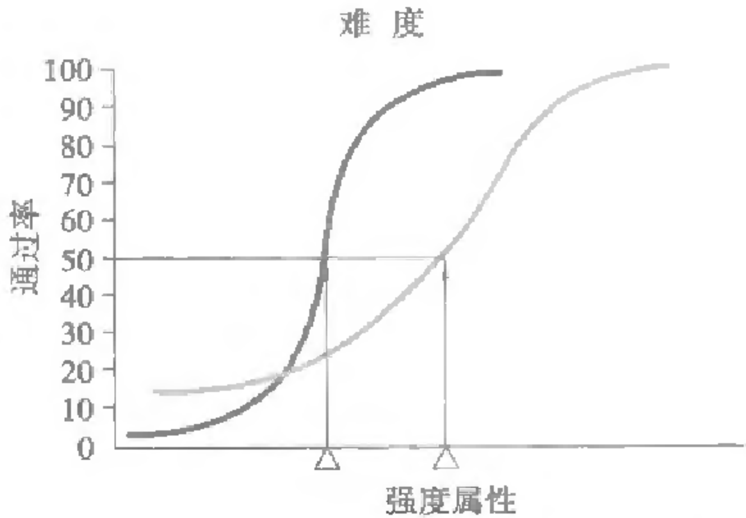


图 7.4 不同难度的两个题项的 ICC 的例子

那就是说，对个体来说，50%的机会通过浅色线所代表的题项比通过深色线所代表的题项要具有更高的属性特征量。使用这个标准，浅色线所代表的题项就更难。既然这样，难度就不是一个主观判断，而是对 X 轴上对应于曲线通过 Y 轴上 0.50 概率值的真实描述。

图 7.5 说明了我们怎样使用同样两条 ICC 曲线来评估区分度。与深色曲线相应的题项在 50%这一通过点比浅色曲线所代表的题项具有更陡的倾斜度。它的结果就是在深色曲线所代表的题项中更小的属性增量就会使明显失败的分数变为明显通过的分数。所以，该题项的更陡的曲线表明，对应的 X 轴上的模糊分数的范围比其他题项的模糊分数的范围更小。这样，在区分失败与通过的人时，深色线所表示题项比浅色线所表示的题项更有效。

最后，在图 7.6 中，我们可以看到，即使当被试的能力(或无论被试要测的什么特征)实际为 0 时，题项曲线仍倾向于向通过分数弯曲。正像你可能猜测的一样，这是由 ICC 曲线与 Y 轴的交点决定的。对深色曲线所表示题项来说，截距接近于零。这样，如果一个人完全缺乏问题中的属性，他或她通过该题项的概率很小。对

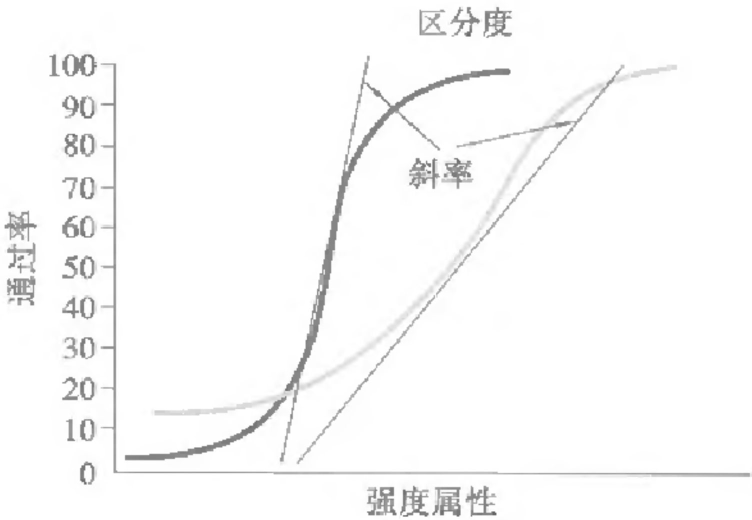


图 7.5　对所测属性具有不同区分度的两个题项的 ICC 的例子

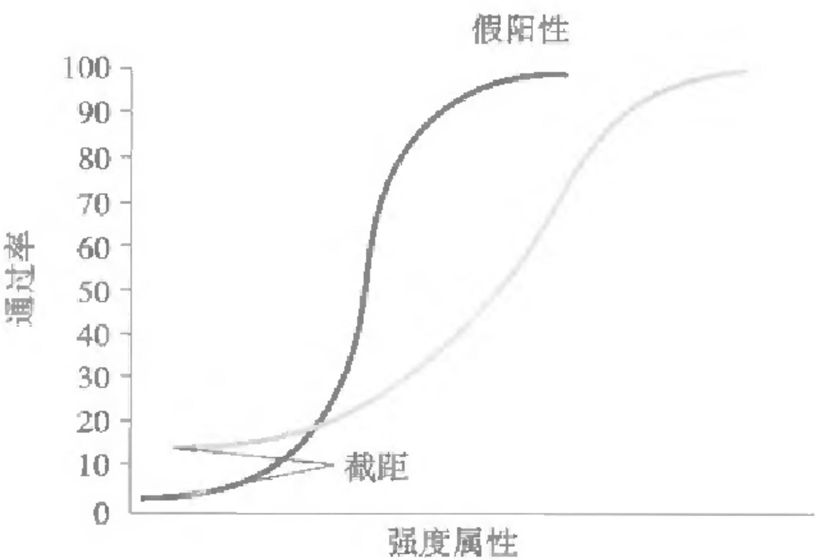


图 7.6　不同假阳性反应率的两个题项的 ICC 的例子

浅色曲线所表示的题项来说，有一个概率(大约 15%)使某些没有能力的人也能通过题项，这样就无法根据问题题项把他们与那些具有很高能力的人相区别。对应的图指出了两个题项在 Y 轴截距的差别——再次得出深色线所表示的题项要更好这个结论。

在理论上，你可以用 IRT 为题项集中的每一个题项建立参数。然后根据计划应用的细节，选出最适宜的题项用来解决手边的测量问题。例如，"简单"题项可以集中起来用来测量那些能力水平相对较低的被试，"难"的题项可以集中起来测量能力水平较高的被试。与此相类似的是，在一个针对儿童的集市或展览会上使用 10 磅的响铃装置而在成人运动员训练营中使用 100 磅的装置。使用不恰当的题项——就像使用不恰当的响铃装置那样——会导致挫折(如果任务太难)或者缺乏积极性(如果任务太简单)。同样，如果编制的测量用来作为一个重要决定的基础，那么缩小每个题项的模糊区间和假阳性的可能性也可能是非常吸引人的。

IRT 方法的突出优势是把我们的注意引向题项表现的三个重要方面(以现在流行的三参数形式)。用源于经典测量理论的方法(例如，通过因素分析或计算 α 系数)，我们可以知道一个题项表现的好还是不好，但是我们不可能对它的缺陷的本质有一个清楚的理解。相比之下，IRT 可以帮我们更明确的评价一个题项的长处和短处。

IRT 的复杂性

虽然 IRT 是十分有吸引力的，但 IRT 不能快速解决测量问题。像传统的测量理论一样，IRT 不能决定题项的特征，只是量化这些特征。就这项技术本身而言，它允许研究者评价题项的绩效，但不能导致一个人直接写出良好的题项或导致较差结构的题项突然工作的良好。此外，当研究者使用基于 IRT 的方法时，评价过程可能会令人气馁。经典的测量通常采用更少有差别且更易处理的概念化的误差源，牺牲精确性来换取简单性。IRT 做了相反的选择，获得精确性但牺牲了简单性。这样，IRT 方法的要求是很高的并且很大程度上被限制在专家的范围内使用。直到 2002 年夏天，

由于没有基于 Windows 的 IRT 分析软件，大多数 IRT 分析软件不得不在没有图形界面、几乎被遗忘的 DOS 操作系统下运行。而且，这些方法的应用还要求具备相当程度的专业判断。这些方法仍然处在一个积极发展的阶段，随着新问题的出现，新的解决办法也将出现。

IRT 的一个主要目标是：要确信对题项特征的评价独立于所研究的样本特征，为了达到这一目的，必须证明这些特征在各个方面，包括能力水平，在不同的样本中是一致的。题项特征不与独立的样本特征，如性别、年龄或其他与测量不相关的变量相联系，这是很重要的。题项分数应该只有在所研究的属性变化时才变化，而不是因为任何其他变量的改变而变化。所以，例如，如果我们假定拼写能力与性别无关，那么我们必须证明具有相同能力的男孩和女孩通过题项的概率相同。如果这不是真的，那么除了拼写能力以外就是性别或其他某种因素在影响题项。同样，对经典理论来说，在一个集合中被检验的题项（如编制一个工具来测量同一变量）必须共享惟一的潜变量。

这些对人和题项的要求指出了另一个棘手的问题：一个工具编制者如何在考虑到 ICC 的情况下确定属性的真实水平（通常称作 θ）。回到我们的木锤和响铃类比，为了确定在一个特殊装置中被试要用多少力量才能敲响铃，你怎样定义强度？在多数情况下，如果属性的真实水平在某种便于管理的形式中是可知的，那么就没有必要再编制一个新的测量工具。在理论上，给出大量人群对一组固定题项的反应，计算机程序应该能够找出题项与个人特征之间的差别。再回到有关狂欢节的类比（响铃装置和浸压机器），如果有足够的人使用每种类型的两个装置，就有可能确定每种类型的哪个样本更难，并且也可判断个体在这两种任务中的技能。在实践中，常常有一个来回迭代的过程，先实施题项来测量特殊被试的属性水平，然后把对该题项的估计作为指标来确定其他题项的特征。当在此基础上确定了最好的题项时，它们就可以在下一轮题项选择中用来获得改良的个体属性水平等等。有些方法依赖于明确的锚定题项（Anchoring Items），这些题项在各种群体中表现稳定，并且可以用作校准其他题项的基础。给出这些过程的本

质，就不难看出为什么 IRT 非常受编制 GRE 考试这样的商业能力测验机构的青睐了。持续的实施和评价为找到那些虽然其他被试特征在大范围的变化但仍能保持特征稳定的题项提供了良好的基础。

何时使用 IRT

有两种特殊的情形使 IRT 方法的优点显得非常重要：

项目等级

首先涉及的是那些本来就是等级的题项。回忆一下，在经典理论中，我们假定题项是潜变量的大致相同的指标。我们假设每个题项对所研究现象的灵敏度大致相等。这些假设在评价许多个人特征时很合适，如态度、信仰和情绪状态。在这时，研究者要测量的变量是一个连续体，并且题项也是据此而构造的。身体能力的测量通常与此形式相一致。例如，一个测量灵活性的测验，用“是”或“否”作为反应选项，可能包含这样的题项，该题项评价一个被试①能独立行走，②只能借助辅助装置行走，或③不能行走。这些题项是不连续的。每个都代表了所研究属性的一个不同的水平。因此，在这种情况下，IRT 测量模型可能比经典模型更合适。

注意，题项本来是连续的，但应用分级反应来进行选择的情况与此不同。例如，一个是两个表达大致相等抑郁程度的题项，一个是带有同意—不同意反应的 6 点量表。既然这样，你可能期望对同一个体来说，所选择的反应选项在所有题项上都保持一致。用我们上面提到的描述灵活性的那种等级题项就不行。实际上，对一个题项回答“是”（如“不能行走”）将会与对另一个回答“是”（如“能独立行走”）不一致。因为每个题项自身变成了属性的一个特殊水平，类似于瑟斯顿或加特曼量表的形式出现，对此，我在第 5 章已做过讨论。

具有等级题项的 IRT 的另一个优点是建立一个由与一个特殊的能力范围相协调的每个题项组成的题项库是可能的。那么我们就可以在身边情形的应用中选择题项。这使测验实施者可以集中

注意于恰当的特征水平，在合适的能力范围内选择题项，减少管理大量域外(Out-of-range)题项的需要。例如，如果某些题项两两间具有连续的等级关系，测验实施者可以选择一个对所测特征的水平要求相对较低的题项，也可以选择另一个要求较高水平的题项。这种特殊选择基于对被试能力水平的最初评估或被试对最初的探测题项的回答情况。如果一个被试通过了简单题项而没有通过困难题项，那么只有难度在两者之间的题项需要进一步考虑。这显然比考虑从最简单到最难的所有题项要方便得多。因为 IRT 能对题项和能力水平之间的匹配进行调整，因而这种管理模式是灵活的(Jenkinson, Fitzpatrick, Garratt, Peto, & Stewart-Brown, 2001)。通常，这种题项管理模式是计算机化，计算机自适应测验(CAT; Van der Linden & Glas, 2000)就是指它。

由于认知能力的显著差异，诸如态度这样的心理变量不像健康状态这样的生理变量那样经常使用 IRT 方法来测量。然而，有些心理变量能很好的符合 IRT 模型。例如，自我效能通常用等级题项来测量，向被试呈现越来越具有挑战性的任务或情境，用来对测量中的轻松或信心进行评价(Devellis & Devellis, 2001)。这似乎是一个 IRT 具有潜在优势的情境。值得注意的是，尽管等级题项的结构和结果与经典测量假设不一致，基于经典模型的对自我效能(和其他变量)的测量似乎运作得相当好。

不同题项的功能

IRT 特别有优势的第二种情形是在区分组群特征和题项特征的差异时。这样的研究注重不同题项的功能，或叫 DIF，也就是说，一个题项在不同的被试群体中会产生不同的反应倾向，而这些被试实际上拥有我们所要评价的同样的特征。例如，如果一个关于抑郁的测验在两个不同年龄组中显示出差异，那可能是因为两个原因中的一个或两个。也就是说，年长的和年轻的人在抑郁上可能存在真实的差异，或者当年长的和年轻的人在抑郁水平相似时他们对特定题项的反应可能存在差异。为了直接比较两个组，我们必须假设在测量时两个组的反应相等并且任何观察到的变异只能归因于所研究的属性。这种假设通常是合理的，当群组在某些方面，如文化或年龄，存在差异，而在题项的解释中这种差异又能

可能合理地导致组间变异时，就要求实验验证。IRT 在这种情况下是一种强大的工具。尽管用 IRT 方法在群体(如种族群体)中收集广泛的数据是回难的，但这些方法为评价那些广泛观察到的差异是由于题项表现还是由于组同差异提供了更高级的工具。经典方法也许能部分完成这一功能(如通过记录群体中明显不同的因素模式)，但可能不会发现更细微的过程。尽管基于经典模型与基于 IRT 模型的结论在特定情境中是否存在分歧是一个经验的问题，但是想象这种情况的出理并不困难。我怀疑 IRT 将越来越多地对不同社会的相关个体进行比较。

值得注意的是，在许多实际问题的评估中有既需要等级题项又需要 DIF 这样的潜在机会。教育评估就是一个(当然不是惟一的一个)已经应用 IRT 的领域。健康结果评估是另一个特别适合的例子。研究的目的常常是分等级的。例如，经过一个医疗程序，一个人经历多少痛苦、伤残或社会孤立是沿连续体变化的，并且个别题项可能对应于那些连续体的不同点。因此，正面的回答两个题项会有不同的意义，这取决于这些题项在问题的连续体中所处的位置。这就需要某种能够鉴别并处理个别不同质的题项的评分系统。IRT 能够解决这个问题。

此外，通过 DIF 区分出真实的组间差异，决策者能敏锐地意识到不同种族群体中的健康差异，并且能够准确的量化它们。IRT 模型似乎又特别适合解决这样的问题。在这个领域，研究者使用基于 IRT 的测量程序已经相当积极。

结　论

基于 IRT 的测量模型有许多引人注目的特征。但不管指导量表编制过程的理论框架是什么，要编制好的题项是艰苦的工作。编制出能一致地测量所研究的属性面对被试的其他特征不敏感的题项是非常有价值的。然而，在基于经典理论的测量中，题项能够在一定程度上弥补其他题项的不足，而 IRT 的逻辑是每个个别题项代表它们自己并且评价自己(尽管编制这样的工具是可能的，其中的题项可能测量同样的现象，但我们前面讨论过，难度可以不

同）。因为一个人能够发现一个表现好的题项，如通过检查 ICC，并不意味着一个人愿意去发现。具有关于被测属性的独立可靠的知识是 IRT 的一个要求，这个要求虽然很难严格满足，但却能够通过对大量异质样本的重测而近似的得到满足。当这一条没有满足时，要说服批评家相信假设已经充分被证明是非常困难的。

我个人的观点是，在经典测量理论的假设可应用的地方，也就是说，在题项被倾向于作为共同潜变量的同等的指标时，经典测量的易于处理性和可实施性使它们被广泛采用。在另一方面，如果研究的问题包含固有的等级反应或注重 DIF，那么基于 IRT 方法的额外的复杂性可能是最好的选择。然而仅仅使用这些方法绝不能保证得到所期望的结果。研究者必须证明：所选择方法的理论假设已经在可接受的限制范围内被满足了，并且测量工具的信度和效度是可以通过实验来验证的。

IRT 使经典方法过时了吗？许多 IRT 的提倡者指出，CMT 和 IRT 有各自的用途。例如，艾布瑞逊和哈希伯格（Embretson & Hershberger，1999）在他们介绍当代测量方法的变化的开头写道："IRT 和 CMT 方法应该结合为一个全面的方法"（P. 252）。最近召开了一个专家座谈会来检查评价健康差异的测量模型，作为其中的一个成员，大卫·塞拉（David Cella，2001）作了如下评述：

> 在测量界内部存在着一种争论的趋势，即争论在普通的测量传统中一种方法的各个方面优于另一种方法的各个方面。随着时间的流逝，与其说给我留下深刻印象的是差异，还不如说我印象最为深刻的是测量方法的共性，或更重要的是由方法而得出的结论。经典测量理论和项目反应理论也许在对个体回答问题的处理方式和记分方式上有明显差异，但是在一种方法下得到的结果很少显著偏离在另一种方法下得到的结果。

IRT 将会更加普及，经典方法也将共存，就像回归分析分享结构方程模型方法的舞台。尽管 IRT 和 SEM 都比它们的前身更有用，但早期的方法仍然保持它们的效用。

8 广泛研究背景下的测量

Measurement in the Broader Research Context

编制量表之前

量表施测之后

最后的思考

这本书开篇就已提供了一些实例来说明测量从何时开始，为何兴起，讨论了在测量中理论的作用并且强调了在测量过程中马马虎虎地实施测量的危害性。在重点转向后面章节的特殊问题之前，概略介绍了研究的广泛背景。本章将在一个更广泛的研究背景中简要地审视量表的应用前景。

编制量表之前

寻求现存的工具

在这本书的前几部分，曾说过量表的编制经常是缺乏适当的已有工具的结果。确定没有适合的、可供选择的测量方法，这是很重要而有效的。在其他地方(DeVellis,1996)我已经建议过寻找适当的量表的方法。通常，这个过程包括寻找印刷的和电子版的测量，从而大概确定是否已经存在合适的量表。一系列的出版物例如心理测量年鉴(the Mental Measurements Yearbook：e. g.，Kramer & Conolwy，1992)和印刷中的测量(Tests in Print：e. g.，Murphy，Conoley，& Impara，1994)包含基础的临床测量，包括能力和人格的测验。经常有些应用心理学家将其用来作为评估来访者的工具。虽然在这些量表中，主要用于做研究的量表较少，但还是包含一些这样的量表。另一种办法是根据汇编而成的书、报、杂志等来寻找合适的量表，例如：人格与社会心理态度测量(Measures of Personality and Social Psychological Attitudes：Robinson，Shaver，& Wrightsman，1991)。在相关的期刊杂志中也可以极佳地发现对相同领域感兴趣的人已经成功运用过的测量策略。

随着测量频率的不断增加，对测量工具的信息的编辑正被放在万维网(World Wide Web)上。实际上，网站是把与测量相关的信息以最快速度扩展的地方。定位于一些特殊研究主题(例如：退伍军人、老年人、少数民族问题)的国际互联网，有时也包括一些在这种类型的研究中使用的测量方法的书目。一些网站特别乐于为确定适当的测量工具提供帮助。

MEI(measurement excellent initiative)是一个特别有用的、可以进入的网页资源,这是老兵事务部门的一个网站。他们的网站既是一个有关测量信息的知识库,也是寻找以印刷和网页为基础的测量工具和信息的一条途径。虽然它的最初目的是有关健康服务研究的测量,但是网站还包括与其他应用相关的测量理论和工具的信息。其网址是www.measurementexperts.org。与MEI站点链接的众多站点之一是健康与心理数据库(HaPI, the health and psychology instruments database;行为测量数据服务,behavioral measurement database services,2000),它还可以在线进入很多大学的图书馆。它包含了大量正在收集的在研究中要使用的工具。在通常情况下,一篇文章中的工具最先出版摘要形式。另外,包括相关心理测量学信息的工具,在随后的应用中其信息也会被包括在内。信息的丰富程度大多依赖于有多少信息提供给HaPI。因此,一些没有被深入描述的测量可以在其他的地方找到。除了有一定的限制之外,它能成为确定潜在相关工具是否有价值的资源。

因此在用任何以网页为基础的信息时,应用者需要考虑信息的出处及其可信度。被大学和政府代理机构(例如:MEI站点)资助的站点和其他机构或者组织(例如:HaPI)建立的站点通常有确切而可信的信息。MEI网站囊括了所有的网站,并提供了那些被认为可信的和负责的网站的链接。然而,在通常情况下,使用这些网页信息时要十分小心。因为网上有大量的"质量低劣的书籍",还有大量的国际互联网站虽然以科学正统的语调和面貌出现,但其包含的内容可能并不符合科学性。你从这本书上学到的技巧将帮助你用更严格的眼光评价任何模式的测量信息,并且帮助你决定被描述的测量是否已被证明有足够的信度和效度。

在所研究的群体环境中审视结构

我们已经讨论过在理论上明确清晰的重要性。评估作为研究者的我们所确定的理论结构与我们计划要研究的人的看法和经验是否一致,这通常是很重要的。群组聚焦分析(参见Krueger & Casery,2000)可以作为一种方法来检验作为所研究的结构的基础

的思想对被试来讲是否有道理。例如,在一项关于归因的研究中,人们被要求沿着诸如"可控性"与"不可控性","适用于特殊场合"与"可适用于大多场合",和"我的一种特征"与"环境或情境的一种特征"这三个维度对各种结果的归因进行解释和说明。归因过程的研究已取得很多成果。沿着这些维度,大多数人能够分析结果,例如,面试完了以后被试得到了一份工作,此时要求被试分析导致结果产生的原因。然而,在某些情况下,这种分析方法可能会导致一些问题。例如,在国外,住在乡下的、没有受过教育并且对这种归因方式不熟悉的老年人,让他们沿着这三条维度对疾病和购物决策进行评价是行不通的。经验表明,他们也许仅仅是无法理解任务,因为让他们以这种方式考虑事情太陌生。要求潜在的研究对象讨论相关概念的群组聚焦分析可以弄清楚这个问题,并且群组群焦分析可以排除注定将导致失败的测量策略。

群组聚焦分析也能够揭示在研究中采用的概念与人们的日常语言之间的关系。一位年轻的母亲不会用与市场专家相同的术语来描述其对一个物品的反应。当儿童在没有具体玩具的情景下进行游戏时,前者可能使用"假装"来描述这种情形,然而一位市场研究人员可以用"不直接接触玩具"进行描述。根据她的语言用法构造的题项(例如:你的孩子花多长时间进行假装游戏,不用任何玩具),比根据专家的语言用法构造的题项(例如:你的孩子在非直接接触玩具的游戏中花多长时间),更可能产生一个适合于测量她对她的孩子怎样与各种各样物品相互作用的感知的工具。

请注意:一些研究者主张只在目标人群中选择愿意接受调查问卷的被试。这是可以接受的而且这可能给参加者一种在研究过程中积极参与的强烈感觉。然而,期望非专业者理解应用到题项结构中的技术问题是不公平的,就像在第5章中讨论过的那样。例如,一位非专业者可能更喜欢用褒义的中性词语来描述一个题项,然而一位有经验的量表编制者会认为在对题项进行回答时,良好的表述不应该产生细微变化,从而避免致使题项无用。如果让参加者感觉到他们积极参与了研究是适当的话,我个人主张用多种方式帮助参加者,使他们感觉到他们积极参与了研究,但是研究者保留决定题项最终表述的权利。如果我们随意创建一个不能精

确地测量他们的看法、情感或态度的情境，那么我们就没有尊重我们研究的参加者。我们仅仅是在浪费他们的时间。

还有其他的方法也能用来决定参加者是否理解提问的目的。例如：简单的向人们，这个问题的意思是什么，或在工具性小型测验中要求被试在形成一个答案时大声报告他们的思考过程，这可能是十分有效的。通常重要的一点是，要理解谁将是被试，并且决定哪种概念表达方式对他们来说是最清楚的。

决定量表施测的模式

研究者能通过多种方式收集数据(例如：Dillman，2000)，他们可以根据被试的偏好来选择相匹配的施测模式。相应地，调查者可以考虑采用比打印的调查问卷要好得多的访谈法。应认识到，企图用打印形式完成的量表与题项同用口头语言表示反应的测量有十分不同的性质。例如：如果父母不得不大声向一位访谈者报告而不是对可选择的反应做标记，那么父母可能会更加不愿意承认他们对孩子有高的期望(企图通过不同于自我施测问卷的模式搜集数据的调查者将可以参考以下作品：Lavrakas，1993 及 Fowler&Mangione，1989)。一般在量表编制过程中严格限制一种新量表的使用方法的施测模式是正确的。一个 G-研究(见第 3 章)可以用来决定量表在施测模式上的通用性(generalizability)。

在其他方法或程序的背景中考虑量表

什么问题或研究程序会超越量表本身？这些问题将如何影响人们对量表的反应？农纳利 (Nunnally，1978，pp. 627～677)把背景因素例如反应形式、疲劳和动机等当作偶然变量(contingent variables)。他还指出它们能够对研究产生三个方面的负面影响：①降低量表的信度；②通过建立可靠的变差来源而不是所研究的结构，从而降低效度；③明确地改变了变量之间的关系，例如：使变量之间出现比实际还要高的相关。作为偶然变量如何起作用的例子，情绪感应(mood induction)和认知情境也可以在市场研究的例子中产生影响。例如，如果市场研究者决定在同一份问卷内包括一个沮丧或自尊量表作为他们的期望量表，那么情绪感应就是一

个问题。涉及这些(和其他)结构的量表经常包含了表达个人自己的消极观点的题项。例如,罗森博格自尊量表(Rosenberg,1965)包含了诸如"我感觉我没有多少值得骄傲的"(也有表达积极自我认同的题项)这样的题项。一位没有注意情绪感应的潜在作用的研究者会在编制一个新的量表时会选择一系列自我评定(Self-critical)的题项。而被试在阅读那些总是对自我消极评价的陈述时可能会产生烦躁不安的状态,并可能导致被试认为无论随后将要感知到的是什么,都不同于他已感知到的(Kihlstrom, Eich, Sandbrand, & Tobias, 2000; Rholes, Riskind, & Lane, 1987)。这可以有农纳利提到的三个不利效果中的任何一个,即出现使情感消极的题项时,期望题项能使情感消极题项的含义有轻微的改变,从而降低这些题项在潜在变量上的变异比例。或者,在一个极端的情况中,期望量表中的一些题项能初步地感知其受到情绪题项的影响,使作为测量父母期望的量表具有多个因子并降低了它的效度。最后,某种程度上,被试的心情影响了他们对期望题项的反应,使这个测验的分数与其他与情绪相关的测量人为地具有高相关。

认知情境是相同的现象中的一个更普遍的例子,即,除了情绪以外,一些相关的结构通过使被试集中注意于某些特殊的题目,从而引起了和上述情境相同的现象。例如,先前提到有关被试的收入,他们家庭的财产值和他们每年花多少钱在不同类别消费物品上等题项的期望量表可以暂时地改变他们对孩子的期望。结果,对量表的反应可能反映了一种短暂的无意识状态。由于情绪的改变,这种认知情境通过影响其清晰地反映父母期望的移变,从而对量表的信度和/或效度产生了不利影响。

量表施测之后

当量表被用来从事真实的研究问题之后,不同的问题就出现了。一个主要的问题是如何分析和解释量表所获得的数据。

分析问题

数据分析中的一个问题是，在不同性质的量表中对变量的不同处理技术的合理性。理论上，这本书极力倡导的方法能使量表能适合广泛多样的数据分析方法。虽然，严格地说，采用利克尔特或语义微分反应模式的题项可以是有顺序的，但有大量经验的人主张对量表采用基于间距(interval-based)的分析方法。然而，社会科学领域里哪种方法最适合哪种类型的数据依然争论激烈，的确，这种情况还将继续下去。决定不同反应选项如何影响对潜在变量的估计，是这个领域的一个积极研究的方向。当然，不同的读者对如何对待测量将有不同的期待。例如，心理学家认为利克尔特量表收集的差异水平数据是有用的，流行病学家却不这样认为。也许最具可操作性的方法是：了解在其感兴趣的领域内流行的观点是什么(并采用流行的观点)。

解释问题

假定研究者已经找到一个合适的策略来分析新编写的量表的数据，如何解释数据依然是一个问题。在这个时刻浮现在头脑中的是，在量表编制过程中没有稳定地建立量表的效度。确定效度是一个不断累积、不断进行的过程。而且，效度实际上是一个量表如何被使用的特征，而不是量表本身的特征。例如，一个抑郁量表在评估抑郁时可能是有效的，但在评估普遍的消极情感时可能就不是有效的。

同样，思考某人的发现也是重要的。尤其是如果出现了违反直觉的或违反理论的结果，研究者必须考虑量表在特殊研究(如果范围不是比较大的话)背景下无效的可能性。它可能是量表的效度受不同人群、情境、施测过程的特殊细节或其他维度的分类的限制。例如，假定对父母的期望进行的心理测量是在相对富裕的人群中编制出来的，那么对于不那么富裕的个体来说，测量的效度可能是不可接受的。根据在一定范围内才能有效应用的量表所得出的任何结论，都应考虑以下方面：①目前应用的情境在多大程度上与它最初有效的情境不同；②对量表的效度进行限制的各种可能性；③这些限制对目前研究的意义。

通用性

虽然上一段就不同人群、情境以及研究的其他方面的通用性提出了警告，但该问题仍然需要进一步强调。得出关于组间存在差异的结论，潜在地混淆了所测现象的差异和工具性能之间的差异。如果我们能够假设后者是微不足道的，那么就可以确定观察到的差异是组间差异。但在许多情况下（比如，对随机选择组和指定分配组儿童完成任务的时间进行比较），我们不能确定工具性能之间的差异是否是微不足道的。在某些情况下，（比如，跨文化地区的人群的比较）我就不能假设出现的差异一定是测量分数的差异。第 7 章所讨论过的 DIF，是心理治疗研究中一个活跃的领域。尽管绝大多数研究者不会将确定工具性能之间的差异作为他们自己努力的核心，但是他们应该意识到工具性能之间存在差异的可能性及其对他们的结论的影响。

最后的思考

测量是社会和行为研究中必不可少的方面。不论研究的其他方面计划和执行得多好，测量可以使一项研究成功或失败。我们假设我们所研究的变量符合我们所采用的估计程序。但通常情况下，初步感兴趣的关系存在于两个或更多无法观察到的变量之间，比如我们可能期望得到某种结果却没有考虑到其他可能的结果。由于我们无法直接测量期望或思考过程，所以我们构造了我们希望能够捕获它们的测量。这些测量在某种意义上是对潜在概念定量的反映。只有当这些反映正确时（比如，工具是有效的），我们所观察到的测量之间的关系才能够反映我们所希望能估计到的、不可观察的结构之间的关系。精细的取样、极好的研究计划以及无可挑剔的执行程序也无法改变这一事实。一位研究者如果不理解测量和他们描绘的变量之间的关系，可以毫不夸张地说，他或她就不会明白自己说的是什么。由此看来，使测量精细化、具体化的努力会通过其获得的益处面得到足够回报。

参考文献
References

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting behavior*. Englewood Cliffs, NJ: Prentice Hall.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Anastasi, A. (1968). *Psychological testing* (3rd ed.). New York: Macmillan.

Asher, H. B. (1983). *Causal modeling* (2nd ed.). Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-003. Beverly Hills, CA: Sage.

Barnette, W. L. (1976). *Readings in psychological tests and measurements* (3rd ed.). Baltimore, MD: Williams & Wilkins.

Behevioral Measurement Database Services. (2000). *Ovid Technologies field guide: Health and Psychological Instruments (HAPI)*. Retrieved May 1, 2002, from http://www.ovid.com/documentation/user/field_guide/disp_fldguide.cfm? db=hapidb.htm

Blalock, S. J., DeVellis, R. F., Brown, G. K., & Wallston, K. A. (1989). Validity of the Center for Epidemiological Studies Depression scale in arthritis populations. *Arthritis and Rheumatism*, *32*, 991-997.

Bohrnstedt, O. W. (1969). A quick method for determining the reliability and validity of multiple-item scales. *American Sociological Review*, *34*, 542-548.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Carmines, E. G., & McIver, J. P. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. W. Bohrnstedt & B. F. Borgatta (Eds.), *Social measurement: Current issues* (pp. 65-115). Beverly Hills, CA: Sage.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Cella, D. (2001, May). Commentary provided at the Resource Centers on Minority Aging Research conference on *Measurement Issues in Health Disparities Research in the U. S.*, San Francisco.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equation causal models. *Applied Psychological Measurement*, *14*, 183-196.

Comrey, A. L. (1973). *A first course in factor analysis*. New York: Academic Press.

Comrey, A. L. (1988). Factor analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*, 754-761.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-063. Beverly Hills, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *Dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cureton, E. E. (1983). *Factor analysis: An applied approach*. Hillsdale, NJ: Lawrence Erlbaum.

Currey, S. S., Callahan, L. F., & DeVellis, R. F. (2002). *Five-item Rheumatology Attitudes Index (RAI): Disadvantages of a single positively worded item*. Unpublished paper, Thurston Arthritis Research Center, University of North Carolina at Chapel Hill.

Czaja, R., & Blair, J. (1996). *Designing surveys: A guice to decisions and procedures*. Thousand Oaks, CA: Pine Forge.

Dale, F., & Chall, J. E. (1948). A formula for predicting readability: Instructions. *Education Research Bulletin*, *27*, 37-54.

DeVellis, B. M., & DaVellis, R. F. (2001). Self-efficacy and health. In A. Baum & T. Revenson (Eds.), *Handbook of Health psychology*. Mahwah, NJ: Lawrence Erlbaum.

DeVellis, R. F. (1996). A consumer's guide to finding, evaluating, and reporting on measurement instruments. *Arthritis Care and Research*, *9*, 239-245.

DeVellis, R. F., Blalock, S. J., Holt, K. D., Renner, B. R., Blanchard, L. W., & Klotz, M. L. (1991). Arthritis patients' reactions to unavoidable social comparisons. *Personality and Social Psychology Bulletin*, *17*, , 392-399.

DeVellis, R. F., & Callahan, L. F. (1993). A brief measure of helplessness: The helplessness subscale of the Rheumatology Attitudes Index. *Journal of Rheumatology*, *20*, 866-869.

DeVellis, R. F., DeVellis, B. M., Blanchard, L. W., Klotz, M. L., Luchok, K., & Voyce, C. (1993). Development and validation of the Parent Health Locus of Control (PHLOC) scales. *Health Education Quarterly*, *20*, 211-225.

DeVellis, R. F., DeVellis, B. M., Revicki, D. A., Lurie, S. J., Runyan, D. K., & Bristol, M. M. (1985). Development and validation of the child improvement locus of control (CILC) scales. *Journal of Social and Clinical Psychology*, *3*, 307-324.

DeVellis, R. F., Holt, K., Renner, B. R., Blalock, S. J., Blanchard, L. W., Cook, H. L., Klotz, M. L., Mikow, V., & Harring, K. (1990). The relationship of social comparison to rheumatoid arthritis symptoms and affect. *Basic and Applied Social Psychology*, *11*, 1-18.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. New York: Russell Sage.

Embretson, S. E., & Hershberger, S. L. (1999). Summary and future of psychometric models in testing. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 243-254). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117-140.

Fink, A. (1995). *The survey kit*. Thousand Oaks, CA: Sage.

Fowler, F. J. (1993). *Survey research methods*. Thousand Oaks, CA: Sage.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage.

Fowler, F. J., & Mangione, T. W. (1989). *Standardized survey interviewing*. Newbury Park, CA: Sage.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading*, *21*, 249.

Ghiselli, B. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: Freeman.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harman, H. H. (1976). *Modern factor analysis*. Chicago: University of Chicago Press.

Hathaway, S. R., & McKinley, J. C. (1967). *Minnesota Multiphasic Personality Inventory: Manual for administration and scoring*. New York: Psychological Corporation.

Hathaway, S. R., & Meehl, P. E. (1951). *An atlas for the clinical use of the MMPI*. Minneapolis: University of Minnesota Press.

Idler, E. L., & Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*, *38*, 21-37.

Jenkinson, C., Fitzpatrick, R., Garratt, A., Peto, V., & Stewart-Brown, S. (2001). Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). *Journal of Neurology, Neurosurgery, and Psychiatry*, *71*, 220-224.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 109-134.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement 20*, 141-151.

Keefe, F. J. (2000). Self-report of pain: Issues and opportunities. In A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 317-337). Mahwah, NJ: Lawrence Erlbaum.

Kelly, J. R., & McGrath, J. B. (1988). *On time and method*. Newbury Park, CA: Sage.

Kihlstrom, J. F., Eich, E., Sandbrand, D., & Tobias, B. A. (2000). Emotion and memory: Implications for self-report. In A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 81-103). Mahwah, NJ: Lawrence Erlbaum.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). San Francisco: Brooks/Cole.

Kramer, J. J., & Conoley, J. C. (1992). *The eleventh mental measurement yearbook*. Lincoln, NE: Boros Institute of Mental Measurements.

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research*. Thousand Oaks, CA: Sage.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Sage Applied Social Research Methods Series, Vol. 7. Thousand Oaks, CA: Sage.

Levenson, H. (1973). Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology*, *41*, 397-404.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis*. Mahwah, NJ: Lawrence Erlbaum.

Long, J. S. (1983). *Confirmatory factor analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-033. Beverly Hills, CA: Sage.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84-99.

Mayer, J. M. (1978). Assessment of depression. In M. P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 4, pp. 358-425). San Francisco: Jossey-Bass.

McDonald, R. P. (1984). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*, 376-390.

Murphy, L. L., Conoley, J. C., & Impara, J. C. (1994). *Tests in print IV*. Lincoln, NE: Boros Institute of Mental Measurements.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston: Allyn & Bacon.

Namboodiri, K. (1984). *Matrix algebra: An introduction*. Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-028. Beverly Hills, CA: Sage.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*, 166-180.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruence in the prediction of attitude change. *Psychological Bulletin*, *62*, 42-55.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Rholes, W. S., Riskind, J. H., & Lane, J. W. (1987). Emotional states and memory biases: Effects of cognitive priming and mood. *Journal of Personality and Social Psychology*, *52*, 91-99.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). *Measures of personality and social psychological attitudes*. San Diego, CA: Academic Press.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rotter, J. B. (1966). Generalized expectancies for internal vs. external control of reinforcement. *Psychological Monographs*, *80* (1, Whole No. 609).

Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.). *The five factor model of personality* (pp. 21-50). New York: Guilford.

Smith, P. H., Earp, J. A., & DeVellis, R. F. (1995). Measuring battering: Development of the Women's Experiences with Battering (WEB) scale. *Women's Health: Research on Gender, Behavior, and Policy*, *1*, 273-288.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *State-trait anxiety inven- tory (STAI) test manual for form X*. Palo Alto, CA: Consulting Psychologists Press.

Strahan, R., & Gerbasi, K. (1972). Short, homogenous version of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *28*, 191-193.

Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*, 414-424.

Van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. St. Paul, MN: Assessment Systems.

Wallston, K. A., Stein, M. J., & Smith, C. A. (1994). Form C of the MHLC Scales: A condition-specific measure of locus of control. *Journal of Personality Assessment*, *63*, 534-553.

Wallston, K. A., Wallston, B. S., & DeVellis, R. (1978). Development and validation of the multidimensional health locus of control (MHLC) scales. *Health Education Monographs*, *6*, 161-170.

Weisberg, H., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey research, polling, and data analysis*. Thousand Oaks, CA: Sage.

Wright, B. D. (1999). Fundemental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.) *The new rules of measurement* (pp. 65-104).

Mahwah, NJ: Lawrence Erlbaum.

Zorzi, M., Priftis, K., & Umilitá, C. (2002). Brain damage: Neglect disrupts the mental number line. *Nature*, *417* (09 May), 138-139.

Zuckerman, M. (1983). The distinction between trait and state scales is not arbitrary: Comment on Allen and Potkay's "On the arbitrary distinction between traits and states." *Journal of Personality and Social Psychology*, *44*, 1083-1086.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442.